



**Agnostic Cys34-albumin adductomics and DNA methylation:
implication of N-acetylcysteine in lung carcinogenesis years
before diagnosis**

Journal:	<i>International Journal of Cancer</i>
Manuscript ID	IJC-19-1731.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Dagnino, Sonia; Imperial College London Faculty of Medicine, School of Public Health Bodinier, Barbara; Imperial College London, School of Public Health, Epidemiology and Biostatistics Grigoryan, Hasmik; University of California Berkeley, School of Public Health Rappaport, Stephen; University of California Berkeley Karimi, Maryam; Imperial College London Guida, Florence; International Agency for Research on Cancer, Genetic Epidemiology Group Polidoro, Silvia; HuGeF Edmands, William ; University of California Berkeley Naccarati, Alessio; HuGeF Fiorito, Giovanni; Università degli Studi di Torino, Department of Medical Sciences; Human Genetics Foundation (HuGeF) - Torino Sacerdote, Carlotta; San Giovanni Battista Hospital, Unit of Cancer Epidemiology Krogh, Vittorio; Fondazione IRCCS Istituto Nazionale dei Tumori, Epidemiology and Prevention Unit Vermeulen, Roel; Utrecht University Vineis, Paolo; Imperial College London; HuGeF Chadeau-Hyam, Marc; Imperial College London
Key Words:	adductomics, Lung Cancer, DNA-methylation, Biomarkers, Oxidative Stress, Smoking

SCHOLARONE™
Manuscripts

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Agnostic Cys34-albumin adductomics and DNA methylation: implication of N-acetylcysteine in lung carcinogenesis years before diagnosis

S. Dagnino^{1,2*}, B. Bodinier^{1,2}, H. Grigoryan³, S.M. Rappaport³, M. Karimi^{1,2}, F. Guida⁴, S. Polidoro⁵, W.B. Edmands³, A. Naccarati⁵, G. Fiorito⁵, C. Sacerdote⁶, V. Krogh⁷, R. Vermeulen^{1,2,8}, P. Vineis^{1,2,5,†}, M. Chadeau-Hyam^{1,2,8†}.

*Corresponding author:
Dr Sonia Dagnino, Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1 PG, UK, sonia.dagnino@imperial.ac.uk, +44 7480 127 942.

- ¹ MRC-PHE Centre for Environment and Health, Imperial College London, London, United Kingdom
² Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom
³ School of Public Health, University of California, Berkeley, California, United States
⁴ International Agency for Research on Cancer (IARC), Lyon, France
⁵ IIGM, Italian Institute for Genomic Medicine, Torino, Italy
⁶ Piedmont Reference Centre for Epidemiology and Cancer Prevention (CPO Piemonte), Italy.
⁷ Fondazione IRCCS - Istituto Nazionale dei Tumori, Milan, Italy.
⁸ Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

† These authors jointly supervised this work.

Keywords: Adductomics; Lung Cancer; DNA-methylation; Biomarkers; Oxidative Stress; Smoking

Novelty & Impact Statements: We detected for the first-time lower levels of N-acetylcysteine (NAC) adduct in lung cancer cases years before diagnosis. This variation was associated to smoking and hypomethylation at certain smoking related CpGs. Our results indicate a perturbation in the oxidative stress pathways in future cases. This provides biological support for the need of further investigation on the role of oxidative stress in carcinogenesis and enquires on the use of NAC as a preventive drug for lung cancer.

Abbreviations

- AUC: Area under the Curve
BMI: Body Mass Index
CSI: Comprehensive Smoking Index
Cys34: Cysteine 34
Cys-Gly: Cysteinyl-Glycine
EPIC: European Prospective Investigation into Cancer and Nutrition
FWER: Family-Wise Error Rate
HPCL: High Pressure Liquid Chromatography
HSA: Human Serum Albumin

HRMS: High Resolution Mass Spectrometry

LOQ: Limit of quantitation

LCC: large-cell carcinomas

NAC: N-acetylcysteine

OS: Oxidative Stress

ROC: Receiver Operating Characteristic

ROS: Reactive Oxygen Species

TtD: Time to Diagnosis

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Abstract (146 words)

Although smoking and oxidative stress are known contributors to lung carcinogenesis, their mechanisms of action remain poorly understood. To shed light into these mechanisms, we applied a novel approach using Cys34-adductomics in a lung cancer nested case-control study (n=212). Adductomics profiles were integrated with DNA-methylation data at established smoking-related CpG sites measured in the same individuals. Our analysis identified 42 Cys34-albumin adducts, of which 2 were significantly differentially abundant in cases and controls: adduct of N-acetylcysteine (NAC, $p=4.15 \times 10^{-3}$) and of Cysteinyl-Glycine (Cys-Gly, $p=7.89 \times 10^{-3}$). Blood levels of the former were found associated to the methylation levels at 11 smoking related CpG sites. We detect, for the first time in prospective blood samples, and irrespective of time-to-diagnosis, decreased levels of NAC adduct in lung cancer cases. Altogether, our results highlight the potential role of these adducts in the oxidative stress response contributing to lung carcinogenesis years before diagnosis

Introduction

Lung cancer is the leading cause of cancer mortality worldwide^{1, 2}, and is causally linked to smoking³⁻⁶. Using high throughput DNA-methylation assays, recent studies have identified epigenetic biomarkers of smoking exposure and smoking history^{4, 7-9}. Some of these markers were also found to be associated with lung cancer risk prospectively¹⁰⁻¹², suggesting the involvement of DNA methylation changes in the mediation of the adverse effect of smoking. However, in the absence of information functionally linking methylation levels at these CpG sites to general biological pathways, these results only provide a fractional insight into the molecular mechanisms involved in (smoking-induced) lung carcinogenesis¹³.

Oxidative stress (OS) has been reported to be related to smoking¹⁴ and is an established contributor to carcinogenesis^{15, 16}, notably through reactive oxygen species (ROS) whose overproduction can activate OS, and ultimately lead to cell damage, carcinogenesis and chemotherapy resistance^{17, 18}.

As highly reactive species, ROS are rapidly degraded and cannot easily be quantified. However, measuring ROS adducts with human serum albumin (HSA) has been shown to be a promising alternative^{19, 20}. HSA is the most abundant protein in serum and the Cysteine 34 site acts as a scavenger of small molecules such as ROS. Recently, an approach based on untargeted high-resolution mass spectrometry (HRMS) has been developed to measure HSA-Cys34 adducts in serum samples, and identified adducts associated with smoking status¹⁹.

In the present work we propose to apply that agnostic adductomics approach to 212 plasma samples from prospective lung cancer cases (median follow-up 7 years) and healthy controls from the Italian component of the European Prospective Investigation into Cancer and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Nutrition (EPIC-Italy) cohort whose blood samples were also used to generate DNA-methylation profiles. We relate adductomic profiles to smoking metrics and disease outcome and integrate these data with DNA-methylation to capture long-term effects of OS on lung carcinogenesis and to explore the role of smoking and its downstream biological consequences in these processes.

For Peer Review

Materials and Methods

Study Population

Plasma samples from study participants were recruited between 1993 and 1998 as part of the Italian branch of the European Prospective Investigation into Cancer and Nutrition Cohort (EPIC)^{21, 22}. Our study population includes 106 prospective lung cancer cases identified through local Cancer registries and 106 healthy controls, which were matched by gender, age, recruiting center and year and season of blood collection) from the centers of Turin and Varese (details are available in SI). This study complies with the Declaration of Helsinki and protocol was approved by the Ethics Committees at the Human Genetics Foundation (IIGM, Turin, Italy).

Exposure data

Detailed information on cigarette smoking was collected at enrollment from questionnaires in the EPIC study and included smoking status and smoking intensity (in cigarettes/day), smoking duration, time since smoking cessation (in former smokers). As previously proposed²³, the Comprehensive Smoking Index (CSI) capturing the dynamics of lifelong smoking exposure and possible smoking cessation was calculated from this data²⁴.

Adductomics measurements

The 212 plasma samples were analysed in duplicate following a previously described protocol based on the identification of third largest tryptic peptide of HSA ("T3", sequence ALVLIAFAQYLQQC34PFEDHVK, mass 2432 Da) containing the Cys34 locus and its potential modifications^{19, 25, 26}. Plasma samples were processed for digestion and dilution and subsequently analysed by Nano Liquid Chromatography and HRMS (NanoHPLC-HRMS) with

an Orbitrap Elite coupled with an Ion-trap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) operated in positive ionisation mode. Spectral data was pre-processed using an in-house R script (see SI for details). Peak areas of T3 adducts were normalized with the corresponding peak area of the “housekeeping” HK peptide (LVNEVTEFAK) to account for differential amounts of HSA. Adducts abundance was expressed as PAR (Peak Area Ratio as defined by adduct peak area/HK peptide peak area x 1000)^{19, 25, 26}.

Signal missing in both duplicates was considered below the limit of quantitation (LOQ) and set to $LOQ/\sqrt{2}$ ¹⁹. If only one measurement was missing, it was imputed in cases and controls separately, using nearest neighbor averaging. From our analysis, a total of 5 adduct features were excluded because they were missing in more than 60% of the samples. While this cut-off was arbitrary, we ensured that we did not induce any bias in our data or exclude any informative or disease-relevant features, by checking that these 5 excluded features were (i) not differentially missing in cases and controls, (ii) missing across batches, and (iii) had poor quality spectral data. All adducts PAR levels were log-transformed.

Annotation of assayed adducts was based on the identification of the monoisotopic mass of the calculated added masses to the T3-peptide, MS² spectra, and also used the library of the adducts constructed in previous work.

Methylation data

DNA-methylation data from the Infinium HumanMethylation450 BeadChip assay was also available in our participants from a previous study¹⁰. Sample preparation and data processing were described elsewhere²⁷ and are detailed in SI. We only considered here the 2,671 CpG sites which were found to be associated to smoking in a recent meta-analysis⁹, of which 2,670 were assayed after probes filtering in our dataset.

Statistical analyses

Adductomics profiling

The association of the measured levels of a given adduct and (i) exposure to tobacco smoke, and (ii) prospective lung cancer status was evaluated using linear mixed models where the adduct level was modelled as the response, and smoking exposure or disease status as explanatory variable. To model the repeated measurement design, the participant ID was included as a random intercept in the model. Similarly, nuisance variation was modelled including random intercepts for technical confounders^{28, 29} (see SI).

Statistical significance of the effect of the variable of interest was evaluated using the p-value of likelihood ratio test comparing the model with and without that variable^{28, 30, 31}.

We corrected for multiple testing using a permutation-based approach estimating the per-test significance level to be applied to control the family wise error rate^{32, 33}. Adductomics profiling was conducted for three different measures of exposure to tobacco smoke (smoking status, pack years, and CSI), and for lung cancer status. For lung cancer analyses, the model was further adjusted on smoking status (current, former or never smoker), and was stratified on the most represented histological subtypes in our study (SI Table 1).

For all disease-related adducts, we regressed, in cases only, the adduct level (as predictor) against the time to diagnosis (TtD) as defined by the time elapsed from blood collection to lung cancer diagnosis, adjusting for the same variables.

OMICs integration

We investigated if and how adducts that were found to be differentially abundant between lung cancer cases and healthy controls were also related to smoking, by integrating adducts levels and DNA methylation levels at the smoking related CpG sites. As detailed in the SI, we

ran a series of linear mixed models for each disease-related adduct and each of the 2,670 smoking-related CpG, setting the methylation level as the variable of interest. These models also accounted for technical variation in the methylation data using a two-step strategy first estimating, and second removing technically-induced shifts in measured methylation levels^{23, 30, 34}. Results from these analyses were visualized as a bi-partite network where edges were selected based on statistical significance of the pairwise associations, correcting for 2,670 tests.

ROC analyses

In order to quantify the disease-relevant information brought about by the disease-related adducts we ran series of unconditional logistic models setting the adduct level as predictors. Results were visualized by means of Receiver Operating Characteristic (ROC) curves, and corresponding Area Under the Curve (AUC). To prevent overfitted results we constructed ROC curves using ($N=5,000$) independent subsamples (see SI). We compared models with (i) only smoking (as measured by packyears and/or principal components from the smoking-related CpG sites), (ii) only the disease-related adducts, (iii) the disease related adducts and smoking metrics.

Results

Study Population and Adductomics measurements

Of the 212 participants originally included in the study, 15 (5 controls and 10 cases) were excluded from the analysis due to missing data. Main characteristics of the resulting 101 controls and 96 cases are summarized in SI Table 1, and show, as expected, similar center, gender, age, and BMI distributions in cases and controls, and an excess of current and former smokers in cases. The TtD ranged from 0.27 to 13.89 years (median 7.03 years).

Our untargeted adductomics approach located a total of 42 T3-peptide modifications in our samples. Of these 42 adduct features, 25 were annotated as part of this work, and 17 remain unknown (see SI Table 2). Coefficients of variation among duplicate measurements ranged from 0.22 and 0.5 with a median of 0.32 (SI Table 3). The differential mass of the measured adducts included truncations and additions and ranged from -65.1159 to 462.2047 Da. Annotations were made on the basis of MS and MSⁿ data (and example of MS and MSⁿ data used for annotations is shown for LC34-Nacetylcysteine in Figure 1). The larger amount of annotated adducts resulted in Cys34 disulfides (Cysteine, homocysteine, cysteinyl glycine, glutathione, N-acetylcysteine etc.). Other annotations included sulfoxidation products, truncations and addition of cyanide, crotonaldehyde and benzaldehyde. The full description of measured adducts is available in SI Table 2.

The distribution of all the adducts levels (expressed as PAR) showed right-skewed distributions, and log-transformation attenuated the asymmetric distribution for all adducts (SI Figure 1).

Association studies

As summarized in SI Table 4, we found a general lower abundance of the assayed T3-peptide adducts in relation to smoking status (SI Figure 2). In particular, for cumulative exposure (pack-years) and CSI, the 42 effect size estimates were negative. We found that the levels of 13 adducts were associated with at least one smoking exposure metric at a nominal significance level of 0.05. Only four of these, LC19 - benzaldehyde, LC20 - unknown, LC34 - N-acetylcysteine, and LC35-cysteinyl glycine, were associated with more than one smoking exposure metric. However, none of these associations survived correction for multiple testing.

Analyses of the lung cancer status indicated that the levels of six T3-peptide adducts were different at a nominal significance level of 0.05 (Figure 2) between cases and controls, and for all six adducts, lower levels were observed in cases (Figure 3 and SI Table 5). Of these, the adduct of N-acetylcysteine (NAC), (LC34, $p=4.15 \times 10^{-3}$, SI Table 5) survived the correction for multiple testing based on our permutation-based per-test significance level ($\alpha'=5.4 \times 10^{-3}$) controlling the FWER below 0.1, and the adduct of dehydrated cysteinyl-glycine (LC33 , $p=7.89 \times 10^{-3}$, Cys-Gly-(-H₂O)) was borderline significant. Upon adjustment for smoking status (Model 2) the association was attenuated for both adducts (Figure 2, SI Table 5), and in particular for the NAC adduct, which, however remained nominally significant ($p<2.25 \times 10^{-2}$). In order to account for histological heterogeneity of lung cancer, we stratified our analyses on the three most common histological subtypes represented in our study: adenocarcinoma ($N=41$ cases), large cell carcinoma (LCC; $N=20$ cases), and squamous cell carcinomas ($N=15$). We did not identify any differentially abundant adducts for squamous cell carcinomas (results not shown). However, we found two T3-peptide adducts with lower levels in adenocarcinomas ($p<0.05$). These included LC19 (adduct of benzaldehyde), which was not identified in the analysis of lung cancer as a whole, and Cys-Gly-(-H₂O), which remained nominally associated to adenocarcinoma status upon adjustment for smoking status (Table 1). Analyses restricted to large-cell carcinomas identified five of the six lung cancer-related T3-peptide adducts as being significantly less abundant in cases ($p<0.05$). All these associations remained nominally significant when adjusting for smoking status. Of the five adducts found to be less abundant in LCC cases, NAC and LC1, despite smaller sample size, exhibited larger differences ($\beta<-1.22$) and stronger associations ($p<3.5 \times 10^{-4}$, Table 1) than those observed in relation to lung cancer (i.e. irrespective of histological subtype).

Plasma levels of the 7 adducts found differentially abundant in lung cancer cases did not support any linear trend in relation to TtD ($p>0.24$, SI Table 6). In particular, the levels of NAC adduct were found particularly stable with TtD for lung cancer as a whole ($\beta<10^{-3}$, $p>0.9$), and for both histological subtypes separately ($p>0.23$). We only identified a significant (and decreasing) trend for adducts plasma levels of Cys-Gly-(-H₂O), LC10 (unknown), and in a lesser extent, of Cys-Gly (SI Table 6) in relation to the TtD in prospective LCC cases.

Integration of DNA methylation data

In order to further explore the role of smoking in the adduct-lung cancer associations, we explored the pairwise associations linking adduct levels of the 6 lung cancer-related T3 adducts to methylation levels at each of the 2,670 assayed CpG sites that were recently established as epigenetics markers of smoking status ³⁵ (Figure 4). From a preliminary principal component analysis, we estimated that 110 effective tests were performed for each adduct. Using a Bonferroni corrected per-test significance for that number of tests, we identified 34 CpG sites that were associated to the adduct level of at least one of the lung cancer-related adducts. Of these 34 CpG sites, 6 were associated with smoking status, cumulative smoking exposure, and CSI in our study population (SI Table 7). The levels of LC1, LC35, and NAC adducts were associated to the methylation levels of 17, 15, and 11 smoking-related CpG sites, respectively (including all the CpG sites associated with smoking in our data at a Bonferroni corrected significance level). Conversely, levels of Cys-Gly-(-H₂O) were not found to be associated to the methylation levels of any of the smoking-related CpG site ($p>0.001$).

ROC analyses

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

We restricted our ROC analyses to the two adducts found to be most differentially abundant between lung cancer cases and controls. For both Cys-Gly-(-H₂O) (Figure 5A) and NAC adducts (Figure 5B), we found that the variation in adduct levels modestly contributed to lung cancer prediction with an average AUC around 0.6. The 76 principal components (PC, explaining more than 80% of the variance of the 2,670 smoking-related CpG sites) yielded an AUC of 0.63, and jointly including the adducts and the 76 PCs from the smoking-related CpG sites did not affect the model performances (mean AUC around 0.63 for both adducts). Models including only cumulative smoking exposure yielded better predictive performances (mean AUC=0.71 [0.68-0.73]) and the inclusion of the adduct levels of either NAC (Figure 5A) or Cys-Gly-(-H₂O) (Figure 5B) slightly improved the performances of the model with an average AUC over 0.74. The model including levels of both adducts and smoking exposure yielded similar performances (mean AUC= 0.74, ROC curves not shown).

Logistic models restricted to LCC (Figure 5C) show better predictive performances of the plasma levels of NAC (mean AUC=0.68) and smoking exposure (mean AUC=0.73). Including both variables in the model further improved its predictive performances and the corresponding mean AUC reached 0.82.

Conversely, analyses of the adenocarcinoma subtype (Figure 5D) suggest, as expected, a moderate predictive ability of smoking exposure (mean AUC=0.66) and of the Cys-Gly-(-H₂O) level (mean AUC=0.61), and a modest performance improvement while considering both variables in the model (mean AUC=0.68).

Discussion

In the present work we use a recently established technique for agnostic Cys34 adductomics¹⁹, in order to discover adducts potentially involved in lung carcinogenesis. Based on 212 biobanked samples of prospective lung cancer cases and matched controls,

we successfully measured 42 adducts. Association studies relating adduct levels to smoking exposure showed a systematically lower abundance of most adducts in relation to smoking exposure. However, none of these reached statistical significance while correcting for multiple testing. Previous analyses have detected ethylene oxide and acrylonitrile in 34 pooled blood sample from (n=158) smokers¹⁹. Consistently with what has been observed in a follow up from that study using individual blood samples²⁶, our study, including individual samples of 67 smokers did not detect ethylene oxide and acrylonitrile. The majority of our results are consistent with a previous study¹⁹ reporting inverse associations with the Cys34 oxidation products (Cys34-Gln crosslink and the sulfinic and sulfonic acids) with smoking. Our results suggest an inverse association of the NAC adduct in relation to smoking exposure, which contradicts results from Grigoryan et al. (2016)¹⁹, but is in line by recent results from Lu et al. (2017)²⁶ considering exposure to smoky coal. This association is consistent with the inverse association we also detect in relation to lung cancer.

Our analysis of lung cancer outcome identified 6 differentially abundant adducts in prospective cases compared to controls ($p < 0.05$). The adduct of NAC showed significantly lower levels according to our permutation-based per-test significance level ensuring a $\text{FWER} < 0.1$. The adduct of Cys-Gly-(-H₂O) was also found to be lower in cases, and only the NAC adduct was also nominally associated with smoking exposure metrics ($p < 0.05$).

The adduct of Cys-Gly-(-H₂O) was the only lung cancer-related adduct that was differentially abundant in adenocarcinoma cases and controls, whereas NAC showed larger and more marked differences between LCC cases and controls ($n = 20$, $\beta = -1.22$, $p = 3.41 \times 10^{-4}$) than in all lung cancer cases ($n = 96$, $\beta = -0.59$, $p < 4.15 \times 10^{-3}$).

Integrative analyses of DNA methylation identified 11 (of the 2,670) smoking-related CpG sites that were associated with the NAC adduct, whereas none of the CpG sites were

associated with plasma levels of the adduct of Cys-Gly-(-H₂O). The smoking related CpG sites associated with the NAC adduct ($p < 0.05$), were all directly associated with adduct levels (except for cpg04095045 which showed an inverse association), suggesting that a reduced adduct level was associated to hypomethylation at these CpG sites. Previous findings have already shown a hypomethylation at these sites in smokers³⁶. These associations could indicate a potential similar smoking-induced pathway resulting in or involving hypomethylation and reduced adduct levels at these specific sites.

Our ROC analyses suggest that both adducts have modest predictive abilities for lung cancer as whole but complement disease-related information from smoking exposure. Analyses restricted to LCC cases showed that plasma levels of the NAC adduct reasonably predicted disease outcome and that predictions were improved by including smoking exposure variables.

The main limitation of our study is the small sample size, which hampers our ability to perform refined sensitivity and/or stratified analyses. The technical challenges and costs of the method have not yet permitted the analysis of a validation set. However, our data arise from 2 different populations (from Torino and Varese recruiting centres), and as an internal validation approach, we analysed both datasets separately. Despite resulting lower sample sizes, our conclusions remained unchanged in both sub-populations (results not shown). Nevertheless, our study is the largest (in terms of the number of samples analysed) performed to date using HRMS-based untargeted adductomics.

NAC is a known antioxidant capable of reducing OS through its ability to scavenge ROS. It is a precursor of L-Cysteine, an essential amino acid involved in the synthesis of reduced glutathione³⁷. The lower abundance of the NAC adduct in cases may reflect the lower bioavailability of NAC in serum suggestive of (potentially smoking-induced) dysregulation of

redox control. Several studies have indicated that elevated levels of ROS and perturbation of redox homeostasis are hallmarks of carcinogenesis^{38, 39}. Therefore, maintaining ROS balance appears important to ensure normal cell growth and development. Studies in mice have indicated the ability of NAC supplementation to hinder the carcinogenicity of cigarette smoke following *in utero* exposure⁴⁰. Supplementation with prenatal NAC was shown to inhibits genomic and postgenomic alterations in lung due to cigarette smoke exposure as a consequence of oxidative stress, hence implying the potential utility of such a supplementation in the at-risk population. In a multicentre intervention study, NAC supplementation had no effect on former or current smoker patients affected with lung, head or neck cancer⁴¹. In another study on healthy smoking volunteers, a supplementation of NAC was associated with a lower level of DNA adducts, micronuclei in mouth floor and urinary excretion of mutagens⁴². Our findings are in line with these results and are the first to report, using prospective samples from an observational study, in-vivo decreased NAC adduct levels in future lung cancer cases, years before onset.

The peptide Cys-Gly is known to be involved, as a precursor, in the glutathione (GSH) pathway. The γ -glutamyl transferase catabolizes conversion of GSH to Cys-Gly, subsequently to glycine and cysteine, which is further used to resynthesize glutathione. The lower level of the Cys-Gly-(-H₂O) adduct in cases, therefore, also points to a perturbation in redox biology, with possible depletion of GSH. Interestingly, our results suggest that these modifications in the Cys-Gly-(-H₂O) adduct are not directly related to smoking exposure or its downstream consequences.

Both our findings are consistent with a perturbation in redox pathways and consequently a potential imbalance in ROS levels and elimination in future lung cancer cases. The

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

perturbation created in the NAC pathway seems to be related to smoking, whereas the effects expressed by the modifications of Cys-Gly do not seem directly related to smoking. Altogether our approach provides interesting new insights into lung carcinogenesis that is consistent with results from previous trials establishing NAC as a potential preventive drug for oxidative stress in smokers and triggers grounds for further investigation of these preventive treatments. Our results highlight the ability of Cys34 adductomics to explore the biological processes by which OS and redox balance dysregulation may contribute, years before clinical manifestations, to carcinogenesis and to investigate the role of smoking in these processes.

For Peer Review

Funding

This work was supported by Horizon 2020 Marie Skłodowska-Curie fellowship EXACT Identifying biomarkers of EXposure leading to Lung Cancer with AdduCTomics (Grant # 708392 to SD). SD acknowledges support from a pump-priming award from the MRC-PHE Centre for Environment and Health (MR/L019744/1). MC-H, BB, RV, and PV acknowledge support from Cancer Research UK Population Research Committee 'Mechanomics' project grant (Grant #22184 to MC-H). PV, MC-H, RV, SR, SD also acknowledge support from the European Commission FP7 framework 'Exposomics' project grant (EU FP7 grant #308610 to PV). SR acknowledges support from grant R33CA191159 from the National Cancer Institute of the U.S. National Institutes of Health.

Acknowledgments

We are thankful to Prof David Phillips and Dr George Preston (and members of their team) for their collaboration.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare no conflict of interests.

Disclaimer: 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.

References

1. IARC. Globocan 2018 : Lung Cancer fact sheet, 2018.
2. WHO. The 10 leading causes of death in the world, 2000 and 2012, vol. 18/01/2015, 2015.
3. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, Chadeau-Hyam M, Ala U, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015;**6**: 10192.
4. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet* 2015;**24**: 2349-59.
5. Vineis P, Alavanja M, Buffler P, Fontham E, Franceschi S, Gao YT, Gupta PC, Hackshaw A, Matos E, Samet J, Sitas F, Smith J, et al. Tobacco and cancer: Recent epidemiological evidence. *J Natl Cancer Inst* 2004;**96**: 99-106.
6. O'Keeffe LM, Taylor G, Huxley RR, Mitchell P, Woodward M, Peters SAE. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ open* 2018;**8**: e021611.
7. Besingi W, Johansson Å. Smoke-related DNA methylation changes in the etiology of human disease. *Hum Mol Genet* 2013.
8. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG Sites Associated with Cigarette Smoking: Analysis of Epigenome-Wide Data from the Sister Study. *Environ Health Perspect* 2014.
9. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, Moreno-Macias H, Smith JA, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* 2016;**9**: 436-47.
10. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, Chung J, Fasanelli F, Guida F, Campanella G, Chadeau-Hyam M, Grankvist K, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International journal of cancer Journal international du cancer* 2017;**140**: 50-61.
11. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, Chadeau-Hyam M, Ala U, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015;**6**.
12. Zhang Y, Breitling LP, Balavarca Y, Holleczer B, Schöttker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. *Int J Cancer* 2016;**139**: 2482-92.
13. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The Exposome: Molecules to Populations. *Annu Rev Pharmacol Toxicol* 2018.
14. De Flora S, Ganchev G, Iltcheva M, La Maestra S, Micale RT, Steele VE, Balansky R. Pharmacological Modulation of Lung Carcinogenesis in Smokers: Preclinical and Clinical Evidence. *Trends in pharmacological sciences* 2016;**37**: 120-42.

15. Kruk J, Aboul-Enein HY. Reactive Oxygen and Nitrogen Species in Carcinogenesis: Implications of Oxidative Stress on the Progression and Development of Several Cancer Types. *Mini reviews in medicinal chemistry* 2017;**17**: 904-19.
16. Prasad S, Gupta SC, Tyagi AK. Reactive oxygen species (ROS) and cancer: Role of antioxidative nutraceuticals. *Cancer letters* 2017;**387**: 95-105.
17. Laurent A, Nicco C, Chereau C, Goulvestre C, Alexandre J, Alves A, Levy E, Goldwasser F, Panis Y, Soubrane O, Weill B, Batteux F. Controlling tumor growth by modulating endogenous production of reactive oxygen species. *Cancer research* 2005;**65**: 948-56.
18. Wang H, Gao Z, Liu X, Agarwal P, Zhao S, Conroy DW, Ji G, Yu J, Jaroniec CP, Liu Z, Lu X, Li X, et al. Targeted production of reactive oxygen species in mitochondria to overcome cancer drug resistance. *Nat Commun* 2018;**9**: 562.
19. Grigoryan H, Edmands W, Lu SS, Yano Y, Regazzoni L, Iavarone AT, Williams ER, Rappaport SM. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Analytical chemistry* 2016;**88**: 10504-12.
20. Rappaport SM, Li H, Grigoryan H, Funk WE, Williams ER. Adductomics: characterizing exposures to reactive electrophiles. *Toxicology letters* 2012;**213**: 83-90.
21. Palli D, Berrino F, Vineis P, Tumino R, Panico S, Masala G, Saieva C, Salvini S, Ceroti M, Pala V, Sieri S, Frasca G, et al. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori* 2003;**89**: 586-93.
22. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997;**26 Suppl 1**: S6-14.
23. Sandanger TM, Nøst TH, Guida F, Rylander C, Campanella G, Muller DC, van Dongen J, Boomsma DI, Johansson M, Vineis P, Vermeulen R, Lund E, et al. DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Scientific Reports* 2018: In Press.
24. Leffondre K, Abrahamowicz M, Xiao Y, Siemiatycki J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Stat Med* 2006;**25**: 4132-46.
25. Liu S, Grigoryan H, Edmands WMB, Dagnino S, Sinharay R, Cullinan P, Collins P, Chung KF, Barratt B, Kelly FJ, Vineis P, Rappaport SM. Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. *Environ Sci Technol* 2018;**52**: 2307-13.
26. Lu SS, Grigoryan H, Edmands WM, Hu W, Iavarone AT, Hubbard A, Rothman N, Vermeulen R, Lan Q, Rappaport SM. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environmental science & technology* 2017;**51**: 46-57.
27. Campanella G, Gunter MJ, Polidoro S, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Fiorito G, Guarrera S, Iacoviello L, Bergdahl IA, et al. Epigenome-wide association study of adiposity and future risk of obesity-related diseases. *Int J Obes* 2018.
28. Chadeau-Hyam M, Vermeulen RC, Hebls DG, Castagne R, Campanella G, Portengen L, Kelly RS, Bergdahl IA, Melin B, Hallmans G, Palli D, Krogh V, et al. Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Ann Oncol* 2014;**25**: 1065-72.
29. McHale CM, Zhang LP, Lan Q, Vermeulen R, Li GL, Hubbard AE, Porter KE, Thomas R, Portier CJ, Shen M, Rappaport SM, Yin SN, et al. Global Gene Expression Profiling of a Population Exposed to a Range of Benzene Levels. *Environ Health Perspect* 2011;**119**: 628-34.

30. Castagne R, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Kleinjans J, de Kok T, Kyrtopoulos SA, et al. Biological marks of early-life socioeconomic experience is detected in the adult inflammatory transcriptome. *Sci Rep* 2016;**6**: 38705.
31. Castagne R, Delpierre C, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Kyrtopoulos S, Hosnijeh FS, et al. A life course approach to explore the biological embedding of socioeconomic position and social mobility through circulating inflammatory markers. *Sci Rep* 2016;**6**: 25170.
32. Chadeau-Hyam M, Ebbels TM, Brown IJ, Chan Q, Stamler J, Huang CC, Daviglus ML, Ueshima H, Zhao L, Holmes E, Nicholson JK, Elliott P, et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *Journal of proteome research* 2010;**9**: 4620-7.
33. Castagne R, Boulange CL, Karaman I, Campanella G, Santos Ferreira DL, Kaluarachchi MR, Lehne B, Moayyeri A, Lewis MR, Spagou K, Dona AC, Evangelos V, et al. Improving Visualization and Interpretation of Metabolome-Wide Association Studies: An Application in a Population-Based Cohort Using Untargeted 1H NMR Metabolic Profiling. *J Proteome Res* 2017;**16**: 3623-33.
34. Vermeulen R, Saberi Hosnijeh F, Bodinier B, Portengen L, Liquet B, Garrido-Manriquez J, Lokhorst H, Bergdahl IA, Kyrtopoulos SA, Johansson AS, Georgiadis P, Melin B, et al. Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses. *Int J Cancer* 2018.
35. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, Moreno-Macias H, Smith JA, et al. Epigenetic Signatures of Cigarette Smoking. *Circulation Cardiovascular genetics* 2016;**9**: 436-47.
36. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics* 2015;**24**: 2349-59.
37. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC cancer* 2012;**12**: 385-.
38. Kawanishi S, Hiraku Y, Pinlaor S, Ma N. Oxidative and nitrative DNA damage in animals and patients with inflammatory diseases in relation to inflammation-related carcinogenesis. *Biological chemistry* 2006;**387**: 365-72.
39. Toyokuni S, Okamoto K, Yodoi J, Hiai H. Persistent oxidative stress in cancer. *FEBS letters* 1995;**358**: 1-3.
40. Balansky R, Ganchev G, Iltcheva M, Steele VE, De Flora S. Transplacental antioxidants inhibit lung tumors in mice exposed to cigarette smoke after birth: a novel preventative strategy? *Current cancer drug targets* 2012;**12**: 164-9.
41. van Zandwijk N, Dalesio O, Pastorino U, de Vries N, van Tinteren H. EUROSCAN, a randomized trial of vitamin A and N-acetylcysteine in patients with head and neck cancer or lung cancer. For the European Organization for Research and Treatment of Cancer Head and Neck and Lung Cancer Cooperative Groups. *Journal of the National Cancer Institute* 2000;**92**: 977-86.
42. Van Schooten FJ, Besaratinia A, De Flora S, D'Agostini F, Izzotti A, Camoirano A, Balm AJ, Dallinga JW, Bast A, Haenen GR, Van't Veer L, Baas P, et al. Effects of oral administration of N-acetyl-L-cysteine: a multi-biomarker study in smokers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for*

Cancer Research, cosponsored by the American Society of Preventive Oncology 2002;**11**: 167-75.

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1: HRMS data acquired for the adduct of N-acetylcysteine. (A) Isotope distribution for the putative N-acetylcysteine adduct in a plasma sample, (B) MS/MS spectrum of the precursor ion at $m/z = 865.43 (+3)$ acquired in the ion trap. The $\Delta m/z$ used to identify the adducted mass is calculated from the shift in the y^{2+} series.

Figure 2: Manhattan plot summarising the results of the association study relating each of the 42 T3 adducts level in relation to lung cancer disease status, using a univariate linear mixed model approach. Each adduct is represented by its $-\log_{10}$ p-value testing the null hypothesis of no difference in adduct level in cases and controls (Y axis) for each assayed T3 adduct, in function of their m/z ratio (X axis). Results are presented for the model adjusting for technical confounders, age, gender and BMI (Model 1, bullets), and for the model additionally adjusting for smoking status (Model 2, triangle). Adducts with p-values above 0.05 are coloured in grey, and those with p-values below 0.05 in dark red. The blue horizontal dashed line represents the nominal 0.05 significance level and the red and green dashed line represent the permutation-based per-test significance level (α') required to control the family wise error rate (FWER) at a 0.1 level for Model 1 and Model 2, respectively. The plain blue line represents the m/z density assayed in our study along the spectrum.

Figure 3: Distribution of the log-transformed level of the 6 T3 adducts associated with Lung Cancer status. Results are presented for healthy controls (light grey) and lung cancer controls (dark grey) separately.

Figure 4: Bi-partite network representation of the pairwise associations between each of the six lung cancer related T3 adducts and the methylation levels at the 2,670 CpG sites found associated to smoking in Joehanes R *et al*, 2016 and assayed in our study population.

Pairwise association was evaluated using a linear mixed model setting the log-transformed adducts levels as outcome and the methylation M-values as predictor. Results are adjusted for technical confounders, age, gender, and BMI. In order to correct for multiple testing, we only represent edges if the p-value for the regression coefficient linking the adduct level and the DNA methylation M-value is below $0.05/110$, where 110 is the number of principal components required to explain more than 90% of the variance of the original (2,670 dimensional) DNA methylation matrix. For clarity, we only represent CpG sites that are found associated to at least one T3 adduct. The CpG sites are coloured according to the p-value measuring their association with smoking status in our study population, and edges are coloured according the direction of the adduct-CpG site association: red for direct associations ($\beta > 0$), and blue for inverse associations ($\beta < 0$).

Figure 5: Receiver Operating Characteristic (ROC) curves from logistic models regressing the disease risk against the adduct levels of NAC (LC34) (left, figures A and C), and Cys-Gly-(H₂O) (LC33, right, figures B and D). Models were fitted using either lung cancer status, irrespective of the histological subtype (top panel, figures A and B), or restricting the cases to large cell carcinomas for NAC adduct (figure C) and to adenocarcinomas for Cys-Gly-(H₂O) adduct (figure D). In each case we ran three different models: one including smoking exposure as measured by pack-years (in red), one including the level of the adduct of interest (in green), one including both the adduct level and the smoking exposure (in light blue). For the analyses of lung cancer (all subtypes) we also investigated a model including the 76 principal components explaining >90% of the variance of the 2,670 smoking-related CpG sites (orange), and the model including both these components and the adduct level (dark blue). We used a subsampling procedure (repeated independently 5,000 times) of 80% of the study population as training set and report the ROC curves and corresponding AUC in

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

the validation set as defined by remaining 20% of the population. The plain ROC curve (and the point estimate of the AUC) correspond to the average performance yielded across the 5,000 subsamples, and the coloured areas (and AUC ranges) reflect the extreme performances yielded across the sub-samples.

For Peer Review

Table 1: Results from the association study of the log-transformed levels of the (N=42) T3 adducts in relation to the two most represented histological subtypes of lung cancer in our study. Results are summarised by the effect size estimate (regression coefficient β) and the corresponding p-value testing the null hypothesis of no association ($\beta=0$). We report results for the model adjusting for technical confounders, age, gender and BMI (Model 1), and for the model additionally adjusting for smoking status (Model 2). For readability, we only report the adducts whose level were found significantly different in cases and controls at a nominal significance level of 0.05 for at least one histological subtype.

	Full sample (96 cases)				Adenocarcinoma (41 cases)				Large-cell carcinoma (20 cases)			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value
LC1 - Unknown	-0.461	4.10e-02	-0.355	1.44e-01	-0.076	6.81e-01	-0.104	5.87e-01	-1.656	3.26e-04	-1.449	2.41e-03
LC10 - Unknown	-0.652	2.18e-02	-0.579	6.01e-02	-0.522	1.13e-01	-0.479	1.61e-01	-1.33	1.36e-02	-1.172	3.84e-02
LC19 - Benzaldehyde	-0.502	1.22e-01	-0.414	2.40e-01	-0.885	4.74e-02	-0.851	6.59e-02	-0.805	1.65e-01	-0.726	2.37e-01
LC26 - Unknown	-0.368	3.19e-02	-0.329	7.58e-02	-0.27	2.16e-01	-0.288	2.05e-01	-0.709	1.98e-02	-0.62	5.14e-02
LC33 - S-Cys-Gly -(H2O)	-0.489	7.89e-03	-0.51	1.07e-02	-0.587	2.22e-02	-0.633	1.77e-02	-0.505	7.50e-02	-0.48	1.09e-01
LC34 - S-(Nacetyl)Cys	-0.587	4.15e-03	-0.505	2.25e-02	-0.229	2.85e-01	-0.154	4.87e-01	-1.219	3.41e-04	-1.199	8.88e-04
LC35 - S-Cys-Gly	-0.315	2.56e-02	-0.302	4.85e-02	-0.069	6.81e-01	-0.119	4.97e-01	-0.554	1.95e-02	-0.627	1.23e-02

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Agnostic Cys34-albumin adductomics and DNA methylation: implication of N-acetylcysteine in lung carcinogenesis years before diagnosis

S. Dagnino^{1,2*}, B. Bodinier^{1,2}, H. Grigoryan³, S.M. Rappaport³, M. Karimi^{1,2}, F. Guida⁴, S. Polidoro⁵, W.B. Edmands³, A. Naccarati⁵, G. Fiorito⁵, C. Sacerdote⁶, V. Krogh⁷, R. Vermeulen^{1,2,8}, P. Vineis^{1,2,5,†}, M. Chadeau-Hyam^{1,2,8†}.

*Corresponding author:
Dr Sonia Dagnino, Epidemiology and Biostatistics, School of Public Health, Imperial College London, Norfolk Place, London W2 1 PG, UK, sonia.dagnino@imperial.ac.uk, +44 7480 127 942.

- ¹ MRC-PHE Centre for Environment and Health, Imperial College London, London, United Kingdom
² Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, United Kingdom
³ School of Public Health, University of California, Berkeley, California, United States
⁴ International Agency for Research on Cancer (IARC), Lyon, France
⁵ IIGM, Italian Institute for Genomic Medicine, Torino, Italy
⁶ Piedmont Reference Centre for Epidemiology and Cancer Prevention (CPO Piemonte), Italy.
⁷ Fondazione IRCCS - Istituto Nazionale dei Tumori, Milan, Italy.
⁸ Division of Environmental Epidemiology, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands

† These authors jointly supervised this work.

Keywords: Adductomics; Lung Cancer; DNA-methylation; Biomarkers; Oxidative Stress; Smoking

Novelty & Impact Statements: We detected for the first-time lower levels of N-acetylcysteine (NAC) adduct in lung cancer cases years before diagnosis. This variation was associated to smoking and hypomethylation at certain smoking related CpGs. Our results indicate a perturbation in the oxidative stress pathways in future cases. This provides biological support for the need of further investigation on the role of oxidative stress in carcinogenesis and enquires on the use of NAC as a preventive drug for lung cancer.

Abbreviations

- AUC: Area under the Curve
BMI: Body Mass Index
CSI: Comprehensive Smoking Index
Cys34: Cysteine 34
Cys-Gly: Cysteinyl-Glycine
EPIC: European Prospective Investigation into Cancer and Nutrition
FWER: Family-Wise Error Rate
HPCL: High Pressure Liquid Chromatography
HSA: Human Serum Albumin

HRMS: High Resolution Mass Spectrometry

LOQ: Limit of quantitation

LCC: large-cell carcinomas

NAC: N-acetylcysteine

OS: Oxidative Stress

ROC: Receiver Operating Characteristic

ROS: Reactive Oxygen Species

TtD: Time to Diagnosis

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Abstract (146 words)

Although smoking and oxidative stress are known contributors to lung carcinogenesis, their mechanisms of action remain poorly understood. To shed light into these mechanisms, we applied a novel approach using Cys34-adductomics in a lung cancer nested case-control study (n=212). Adductomics profiles were integrated with DNA-methylation data at established smoking-related CpG sites measured in the same individuals. Our analysis identified 42 Cys34-albumin adducts, of which 2 were significantly differentially abundant in cases and controls: adduct of N-acetylcysteine (NAC, $p=4.15 \times 10^{-3}$) and of Cysteinyl-Glycine (Cys-Gly, $p=7.89 \times 10^{-3}$). Blood levels of the former were found associated to the methylation levels at 11 smoking related CpG sites. We detect, for the first time in prospective blood samples, and irrespective of time-to-diagnosis, decreased levels of NAC adduct in lung cancer cases. Altogether, our results highlight the potential role of these adducts in the oxidative stress response contributing to lung carcinogenesis years before diagnosis

Introduction

Lung cancer is the leading cause of cancer mortality worldwide^{1, 2}, and is causally linked to smoking³⁻⁶. Using high throughput DNA-methylation assays, recent studies have identified epigenetic biomarkers of smoking exposure and smoking history^{4, 7-9}. Some of these markers were also found to be associated with lung cancer risk prospectively¹⁰⁻¹², suggesting the involvement of DNA methylation changes in the mediation of the adverse effect of smoking. However, in the absence of information functionally linking methylation levels at these CpG sites to general biological pathways, these results only provide a fractional insight into the molecular mechanisms involved in (smoking-induced) lung carcinogenesis¹³.

Oxidative stress (OS) has been reported to be related to smoking¹⁴ and is an established contributor to carcinogenesis^{15, 16}, notably through reactive oxygen species (ROS) whose overproduction can activate OS, and ultimately lead to cell damage, carcinogenesis and chemotherapy resistance^{17, 18}.

As highly reactive species, ROS are rapidly degraded and cannot easily be quantified. However, measuring ROS adducts with human serum albumin (HSA) has been shown to be a promising alternative^{19, 20}. HSA is the most abundant protein in serum and the Cysteine 34 site acts as a scavenger of small molecules such as ROS. Recently, an approach based on untargeted high-resolution mass spectrometry (HRMS) has been developed to measure HSA-Cys34 adducts in serum samples, and identified adducts associated with smoking status¹⁹.

In the present work we propose to apply that agnostic adductomics approach to 212 plasma samples from prospective lung cancer cases (median follow-up 7 years) and healthy controls from the Italian component of the European Prospective Investigation into Cancer and

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Nutrition (EPIC-Italy) cohort whose blood samples were also used to generate DNA-methylation profiles. We relate adductomic profiles to smoking metrics and disease outcome and integrate these data with DNA-methylation to capture long-term effects of OS on lung carcinogenesis and to explore the role of smoking and its downstream biological consequences in these processes.

For Peer Review

Materials and Methods

Study Population

Plasma samples from study participants were recruited between 1993 and 1998 as part of the Italian branch of the European Prospective Investigation into Cancer and Nutrition Cohort (EPIC)^{21, 22}. Our study population includes 106 prospective lung cancer cases identified through local Cancer registries and 106 healthy controls, which were matched by gender, age, recruiting center and year and season of blood collection) from the centers of Turin and Varese (details are available in SI). This study complies with the Declaration of Helsinki and protocol was approved by the Ethics Committees at the Human Genetics Foundation (IIGM, Turin, Italy).

Exposure data

Detailed information on cigarette smoking was collected at enrollment from questionnaires in the EPIC study and included smoking status and smoking intensity (in cigarettes/day), smoking duration, time since smoking cessation (in former smokers). As previously proposed²³, the Comprehensive Smoking Index (CSI) capturing the dynamics of lifelong smoking exposure and possible smoking cessation was calculated from this data²⁴.

Adductomics measurements

The 212 plasma samples were analysed in duplicate following a previously described protocol based on the identification of third largest tryptic peptide of HSA ("T3", sequence ALVLIAFAQYLQQC34PFEDHVK, mass 2432 Da) containing the Cys34 locus and its potential modifications^{19, 25, 26}. Plasma samples were processed for digestion and dilution and subsequently analysed by Nano Liquid Chromatography and HRMS (NanoHPLC-HRMS) with

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

an Orbitrap Elite coupled with an Ion-trap mass spectrometer (Thermo Fisher Scientific, Waltham, MA) operated in positive ionisation mode. Spectral data was pre-processed using an in-house R script (see SI for details). Peak areas of T3 adducts were normalized with the corresponding peak area of the “housekeeping” HK peptide (LVNEVTEFAK) to account for differential amounts of HSA. Adducts abundance was expressed as PAR (Peak Area Ratio as defined by adduct peak area/HK peptide peak area x 1000)^{19, 25, 26}.

Signal missing in both duplicates was considered below the limit of quantitation (LOQ) and set to $LOQ/\sqrt{2}$ ¹⁹. If only one measurement was missing, it was imputed in cases and controls separately, using nearest neighbor averaging. From our analysis, a total of 5 adduct features were excluded because they were missing in more than 60% of the samples. While this cut-off was arbitrary, we ensured that we did not induce any bias in our data or exclude any informative or disease-relevant features, by checking that these 5 excluded features were (i) not differentially missing in cases and controls, (ii) missing across batches, and (iii) had poor quality spectral data. All adducts PAR levels were log-transformed.

Annotation of assayed adducts was based on the identification of the monoisotopic mass of the calculated added masses to the T3-peptide, MS² spectra, and also used the library of the adducts constructed in previous work.

Methylation data

DNA-methylation data from the Infinium HumanMethylation450 BeadChip assay was also available in our participants from a previous study¹⁰. Sample preparation and data processing were described elsewhere²⁷ and are detailed in SI. We only considered here the 2,671 CpG sites which were found to be associated to smoking in a recent meta-analysis⁹, of which 2,670 were assayed after probes filtering in our dataset.

Statistical analyses

Adductomics profiling

The association of the measured levels of a given adduct and (i) exposure to tobacco smoke, and (ii) prospective lung cancer status was evaluated using linear mixed models where the adduct level was modelled as the response, and smoking exposure or disease status as explanatory variable. To model the repeated measurement design, the participant ID was included as a random intercept in the model. Similarly, nuisance variation was modelled including random intercepts for technical confounders^{28, 29} (see SI).

Statistical significance of the effect of the variable of interest was evaluated using the p-value of likelihood ratio test comparing the model with and without that variable^{28, 30, 31}.

We corrected for multiple testing using a permutation-based approach estimating the per-test significance level to be applied to control the family wise error rate^{32, 33}. Adductomics profiling was conducted for three different measures of exposure to tobacco smoke (smoking status, pack years, and CSI), and for lung cancer status. For lung cancer analyses, the model was further adjusted on smoking status (current, former or never smoker), and was stratified on the most represented histological subtypes in our study (SI Table 1).

For all disease-related adducts, we regressed, in cases only, the adduct level (as predictor) against the time to diagnosis (TtD) as defined by the time elapsed from blood collection to lung cancer diagnosis, adjusting for the same variables.

OMICs integration

We investigated if and how adducts that were found to be differentially abundant between lung cancer cases and healthy controls were also related to smoking, by integrating adducts levels and DNA methylation levels at the smoking related CpG sites. As detailed in the SI, we

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

ran a series of linear mixed models for each disease-related adduct and each of the 2,670 smoking-related CpG, setting the methylation level as the variable of interest. These models also accounted for technical variation in the methylation data using a two-step strategy first estimating, and second removing technically-induced shifts in measured methylation levels^{23, 30, 34}. Results from these analyses were visualized as a bi-partite network where edges were selected based on statistical significance of the pairwise associations, correcting for 2,670 tests.

ROC analyses

In order to quantify the disease-relevant information brought about by the disease-related adducts we ran series of unconditional logistic models setting the adduct level as predictors. Results were visualized by means of Receiver Operating Characteristic (ROC) curves, and corresponding Area Under the Curve (AUC). To prevent overfitted results we constructed ROC curves using (*N*=5,000) independent subsamples (see SI). We compared models with (i) only smoking (as measured by packyears and/or principal components from the smoking-related CpG sites), (ii) only the disease-related adducts, (iii) the disease related adducts and smoking metrics.

Results

Study Population and Adductomics measurements

Of the 212 participants originally included in the study, 15 (5 controls and 10 cases) were excluded from the analysis due to missing data. Main characteristics of the resulting 101 controls and 96 cases are summarized in SI Table 1, and show, as expected, similar center, gender, age, and BMI distributions in cases and controls, and an excess of current and former smokers in cases. The TtD ranged from 0.27 to 13.89 years (median 7.03 years).

Our untargeted adductomics approach located a total of 42 T3-peptide modifications in our samples. Of these 42 adduct features, 25 were annotated as part of this work, and 17 remain unknown (see SI Table 2). Coefficients of variation among duplicate measurements ranged from 0.22 and 0.5 with a median of 0.32 (SI Table 3). The differential mass of the measured adducts included truncations and additions and ranged from -65.1159 to 462.2047 Da. Annotations were made on the basis of MS and MSⁿ data (and example of MS and MSⁿ data used for annotations is shown for LC34-Nacetylcysteine in Figure 1). The larger amount of annotated adducts resulted in Cys34 disulfides (Cysteine, homocysteine, cysteinyl glycine, glutathione, N-acetylcysteine etc.). Other annotations included sulfoxidation products, truncations and addition of cyanide, crotonaldehyde and benzaldehyde. The full description of measured adducts is available in SI Table 2.

The distribution of all the adducts levels (expressed as PAR) showed right-skewed distributions, and log-transformation attenuated the asymmetric distribution for all adducts (SI Figure 1).

Association studies

As summarized in SI Table 4, we found a general lower abundance of the assayed T3-peptide adducts in relation to smoking status (SI Figure 2). In particular, for cumulative exposure (pack-years) and CSI, the 42 effect size estimates were negative. We found that the levels of 13 adducts were associated with at least one smoking exposure metric at a nominal significance level of 0.05. Only four of these, LC19 - benzaldehyde, LC20 - unknown, LC34 - N-acetylcysteine, and LC35-cysteinyl glycine, were associated with more than one smoking exposure metric. However, none of these associations survived correction for multiple testing.

Analyses of the lung cancer status indicated that the levels of six T3-peptide adducts were different at a nominal significance level of 0.05 (Figure 2) between cases and controls, and for all six adducts, lower levels were observed in cases (Figure 3 and SI Table 5). Of these, the adduct of N-acetylcysteine (NAC), (LC34, $p=4.15 \times 10^{-3}$, SI Table 5) survived the correction for multiple testing based on our permutation-based per-test significance level ($\alpha'=5.4 \times 10^{-3}$) controlling the FWER below 0.1, and the adduct of dehydrated cysteinyl-glycine (LC33, $p=7.89 \times 10^{-3}$, Cys-Gly-(-H₂O)) was borderline significant. Upon adjustment for smoking status (Model 2) the association was attenuated for both adducts (Figure 2, SI Table 5), and in particular for the NAC adduct, which, however remained nominally significant ($p < 2.25 \times 10^{-2}$). In order to account for histological heterogeneity of lung cancer, we stratified our analyses on the three most common histological subtypes represented in our study: adenocarcinoma ($N=41$ cases), large cell carcinoma (LCC; $N=20$ cases), and squamous cell carcinomas ($N=15$). We did not identify any differentially abundant adducts for squamous cell carcinomas (results not shown). However, we found two T3-peptide adducts with lower levels in adenocarcinomas ($p < 0.05$). These included LC19 (adduct of benzaldehyde), which was not identified in the analysis of lung cancer as a whole, and Cys-Gly-(-H₂O), which remained nominally associated to adenocarcinoma status upon adjustment for smoking status (Table 1). Analyses restricted to large-cell carcinomas identified five of the six lung cancer-related T3-peptide adducts as being significantly less abundant in cases ($p < 0.05$). All these associations remained nominally significant when adjusting for smoking status. Of the five adducts found to be less abundant in LCC cases, NAC and LC1, despite smaller sample size, exhibited larger differences ($\beta < -1.22$) and stronger associations ($p < 3.5 \times 10^{-4}$, Table 1) than those observed in relation to lung cancer (i.e. irrespective of histological subtype).

Plasma levels of the 7 adducts found differentially abundant in lung cancer cases did not support any linear trend in relation to TtD ($p>0.24$, SI Table 6). In particular, the levels of NAC adduct were found particularly stable with TtD for lung cancer as a whole ($\beta<10^{-3}$, $p>0.9$), and for both histological subtypes separately ($p>0.23$). We only identified a significant (and decreasing) trend for adducts plasma levels of Cys-Gly-(-H₂O), LC10 (unknown), and in a lesser extent, of Cys-Gly (SI Table 6) in relation to the TtD in prospective LCC cases.

Integration of DNA methylation data

In order to further explore the role of smoking in the adduct-lung cancer associations, we explored the pairwise associations linking adduct levels of the 6 lung cancer-related T3 adducts to methylation levels at each of the 2,670 assayed CpG sites that were recently established as epigenetics markers of smoking status ³⁵ (Figure 4). From a preliminary principal component analysis, we estimated that 110 effective tests were performed for each adduct. Using a Bonferroni corrected per-test significance for that number of tests, we identified 34 CpG sites that were associated to the adduct level of at least one of the lung cancer-related adducts. Of these 34 CpG sites, 6 were associated with smoking status, cumulative smoking exposure, and CSI in our study population (SI Table 7). The levels of LC1, LC35, and NAC adducts were associated to the methylation levels of 17, 15, and 11 smoking-related CpG sites, respectively (including all the CpG sites associated with smoking in our data at a Bonferroni corrected significance level). Conversely, levels of Cys-Gly-(-H₂O) were not found to be associated to the methylation levels of any of the smoking-related CpG site ($p>0.001$).

ROC analyses

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

We restricted our ROC analyses to the two adducts found to be most differentially abundant between lung cancer cases and controls. For both Cys-Gly-(-H₂O) (Figure 5A) and NAC adducts (Figure 5B), we found that the variation in adduct levels modestly contributed to lung cancer prediction with an average AUC around 0.6. The 76 principal components (PC, explaining more than 80% of the variance of the 2,670 smoking-related CpG sites) yielded an AUC of 0.63, and jointly including the adducts and the 76 PCs from the smoking-related CpG sites did not affect the model performances (mean AUC around 0.63 for both adducts). Models including only cumulative smoking exposure yielded better predictive performances (mean AUC=0.71 [0.68-0.73]) and the inclusion of the adduct levels of either NAC (Figure 5A) or Cys-Gly-(-H₂O) (Figure 5B) slightly improved the performances of the model with an average AUC over 0.74. The model including levels of both adducts and smoking exposure yielded similar performances (mean AUC= 0.74, ROC curves not shown).

Logistic models restricted to LCC (Figure 5C) show better predictive performances of the plasma levels of NAC (mean AUC=0.68) and smoking exposure (mean AUC=0.73). Including both variables in the model further improved its predictive performances and the corresponding mean AUC reached 0.82.

Conversely, analyses of the adenocarcinoma subtype (Figure 5D) suggest, as expected, a moderate predictive ability of smoking exposure (mean AUC=0.66) and of the Cys-Gly-(-H₂O) level (mean AUC=0.61), and a modest performance improvement while considering both variables in the model (mean AUC=0.68).

Discussion

In the present work we use a recently established technique for agnostic Cys34 adductomics¹⁹, in order to discover adducts potentially involved in lung carcinogenesis. Based on 212 biobanked samples of prospective lung cancer cases and matched controls,

we successfully measured 42 adducts. Association studies relating adduct levels to smoking exposure showed a systematically lower abundance of most adducts in relation to smoking exposure. However, none of these reached statistical significance while correcting for multiple testing. Previous analyses have detected ethylene oxide and acrylonitrile in 34 pooled blood sample from (n=158) smokers¹⁹. Consistently with what has been observed in a follow up from that study using individual blood samples²⁶, our study, including individual samples of 67 smokers did not detect ethylene oxide and acrylonitrile. The majority of our results are consistent with a previous study¹⁹ reporting inverse associations with the Cys34 oxidation products (Cys34-Gln crosslink and the sulfinic and sulfonic acids) with smoking. Our results suggest an inverse association of the NAC adduct in relation to smoking exposure, which contradicts results from Grigoryan et al. (2016)¹⁹, but is in line by recent results from Lu et al. (2017)²⁶ considering exposure to smoky coal. This association is consistent with the inverse association we also detect in relation to lung cancer.

Our analysis of lung cancer outcome identified 6 differentially abundant adducts in prospective cases compared to controls ($p < 0.05$). The adduct of NAC showed significantly lower levels according to our permutation-based per-test significance level ensuring a $\text{FWER} < 0.1$. The adduct of Cys-Gly-(-H₂O) was also found to be lower in cases, and only the NAC adduct was also nominally associated with smoking exposure metrics ($p < 0.05$).

The adduct of Cys-Gly-(-H₂O) was the only lung cancer-related adduct that was differentially abundant in adenocarcinoma cases and controls, whereas NAC showed larger and more marked differences between LCC cases and controls ($n = 20$, $\beta = -1.22$, $p = 3.41 \times 10^{-4}$) than in all lung cancer cases ($n = 96$, $\beta = -0.59$, $p < 4.15 \times 10^{-3}$).

Integrative analyses of DNA methylation identified 11 (of the 2,670) smoking-related CpG sites that were associated with the NAC adduct, whereas none of the CpG sites were

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

associated with plasma levels of the adduct of Cys-Gly-(-H₂O). The smoking related CpG sites associated with the NAC adduct ($p<0.05$), were all directly associated with adduct levels (except for cpg04095045 which showed an inverse association), suggesting that a reduced adduct level was associated to hypomethylation at these CpG sites. Previous findings have already shown a hypomethylation at these sites in smokers³⁶. These associations could indicate a potential similar smoking-induced pathway resulting in or involving hypomethylation and reduced adduct levels at these specific sites.

Our ROC analyses suggest that both adducts have modest predictive abilities for lung cancer as whole but complement disease-related information from smoking exposure. Analyses restricted to LCC cases showed that plasma levels of the NAC adduct reasonably predicted disease outcome and that predictions were improved by including smoking exposure variables.

The main limitation of our study is the small sample size, which hampers our ability to perform refined sensitivity and/or stratified analyses. The technical challenges and costs of the method have not yet permitted the analysis of a validation set. However, our data arise from 2 different populations (from Torino and Varese recruiting centres), and as an internal validation approach, we analysed both datasets separately. Despite resulting lower sample sizes, our conclusions remained unchanged in both sub-populations (results not shown). Nevertheless, our study is the largest (in terms of the number of samples analysed) performed to date using HRMS-based untargeted adductomics.

NAC is a known antioxidant capable of reducing OS through its ability to scavenge ROS. It is a precursor of L-Cysteine, an essential amino acid involved in the synthesis of reduced glutathione³⁷. The lower abundance of the NAC adduct in cases may reflect the lower bioavailability of NAC in serum suggestive of (potentially smoking-induced) dysregulation of

1
2
3 redox control. Several studies have indicated that elevated levels of ROS and perturbation of
4
5 redox homeostasis are hallmarks of carcinogenesis^{38, 39}. Therefore, maintaining ROS
6
7 balance appears important to ensure normal cell growth and development. Studies in mice
8
9 have indicated the ability of NAC supplementation to hinder the carcinogenicity of cigarette
10
11 smoke following *in utero* exposure⁴⁰. Supplementation with prenatal NAC was shown to
12
13 inhibits genomic and postgenomic alterations in lung due to cigarette smoke exposure as a
14
15 consequence of oxidative stress, hence implying the potential utility of such a
16
17 supplementation in the at-risk population. In a multicentre intervention study, NAC
18
19 supplementation had no effect on former or current smoker patients affected with lung,
20
21 head or neck cancer⁴¹. In another study on healthy smoking volunteers, a supplementation
22
23 of NAC was associated with a lower level of DNA adducts, micronuclei in mouth floor and
24
25 urinary excretion of mutagens⁴². Our findings are in line with these results and are the first
26
27 to report, using prospective samples from an observational study, in-vivo decreased NAC
28
29 adduct levels in future lung cancer cases, years before onset.
30
31
32
33
34
35

36
37 The peptide Cys-Gly is known to be involved, as a precursor, in the glutathione (GSH)
38
39 pathway. The γ -glutamyl transferase catabolizes conversion of GSH to Cys-Gly, subsequently
40
41 to glycine and cysteine, which is further used to resynthesize glutathione. The lower level of
42
43 the Cys-Gly-(-H₂O) adduct in cases, therefore, also points to a perturbation in redox biology,
44
45 with possible depletion of GSH. Interestingly, our results suggest that these modifications in
46
47 the Cys-Gly-(-H₂O) adduct are not directly related to smoking exposure or its downstream
48
49 consequences.
50
51
52
53

54 Both our findings are consistent with a perturbation in redox pathways and consequently a
55
56 potential imbalance in ROS levels and elimination in future lung cancer cases. The
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

perturbation created in the NAC pathway seems to be related to smoking, whereas the effects expressed by the modifications of Cys-Gly do not seem directly related to smoking. Altogether our approach provides interesting new insights into lung carcinogenesis that is consistent with results from previous trials establishing NAC as a potential preventive drug for oxidative stress in smokers and triggers grounds for further investigation of these preventive treatments. Our results highlight the ability of Cys34 adductomics to explore the biological processes by which OS and redox balance dysregulation may contribute, years before clinical manifestations, to carcinogenesis and to investigate the role of smoking in these processes.

For Peer Review

Funding

This work was supported by Horizon 2020 Marie Skłodowska-Curie fellowship EXACT Identifying biomarkers of EXposure leading to Lung Cancer with AdduCTomics (Grant # 708392 to SD). SD acknowledges support from a pump-priming award from the MRC-PHE Centre for Environment and Health (MR/L019744/1). MC-H, BB, RV, and PV acknowledge support from Cancer Research UK Population Research Committee 'Mechanomics' project grant (Grant #22184 to MC-H). PV, MC-H, RV, SR, SD also acknowledge support from the European Commission FP7 framework 'Exposomics' project grant (EU FP7 grant #308610 to PV). SR acknowledges support from grant R33CA191159 from the National Cancer Institute of the U.S. National Institutes of Health.

Acknowledgments

We are thankful to Prof David Phillips and Dr George Preston (and members of their team) for their collaboration.

Data availability statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests: The authors declare no conflict of interests.

Disclaimer: 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.

References

1. IARC. Globocan 2018 : Lung Cancer fact sheet, 2018.
2. WHO. The 10 leading causes of death in the world, 2000 and 2012, vol. 18/01/2015, 2015.
3. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, Chadeau-Hyam M, Ala U, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015;**6**: 10192.
4. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet* 2015;**24**: 2349-59.
5. Vineis P, Alavanja M, Buffler P, Fontham E, Franceschi S, Gao YT, Gupta PC, Hackshaw A, Matos E, Samet J, Sitas F, Smith J, et al. Tobacco and cancer: Recent epidemiological evidence. *J Natl Cancer Inst* 2004;**96**: 99-106.
6. O'Keeffe LM, Taylor G, Huxley RR, Mitchell P, Woodward M, Peters SAE. Smoking as a risk factor for lung cancer in women and men: a systematic review and meta-analysis. *BMJ open* 2018;**8**: e021611.
7. Besingi W, Johansson Å. Smoke-related DNA methylation changes in the etiology of human disease. *Hum Mol Genet* 2013.
8. Harlid S, Xu Z, Panduri V, Sandler DP, Taylor JA. CpG Sites Associated with Cigarette Smoking: Analysis of Epigenome-Wide Data from the Sister Study. *Environ Health Perspect* 2014.
9. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, Moreno-Macias H, Smith JA, et al. Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc Genet* 2016;**9**: 436-47.
10. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, Chung J, Fasanelli F, Guida F, Campanella G, Chadeau-Hyam M, Grankvist K, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. *International journal of cancer Journal international du cancer* 2017;**140**: 50-61.
11. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, Chadeau-Hyam M, Ala U, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat Commun* 2015;**6**.
12. Zhang Y, Breitling LP, Balavarca Y, Holleczeck B, Schöttker B, Brenner H. Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality. *Int J Cancer* 2016;**139**: 2482-92.
13. Niedzwiecki MM, Walker DI, Vermeulen R, Chadeau-Hyam M, Jones DP, Miller GW. The Exposome: Molecules to Populations. *Annu Rev Pharmacol Toxicol* 2018.
14. De Flora S, Ganchev G, Iltcheva M, La Maestra S, Micale RT, Steele VE, Balansky R. Pharmacological Modulation of Lung Carcinogenesis in Smokers: Preclinical and Clinical Evidence. *Trends in pharmacological sciences* 2016;**37**: 120-42.

15. Kruk J, Aboul-Enein HY. Reactive Oxygen and Nitrogen Species in Carcinogenesis: Implications of Oxidative Stress on the Progression and Development of Several Cancer Types. *Mini reviews in medicinal chemistry* 2017;**17**: 904-19.
16. Prasad S, Gupta SC, Tyagi AK. Reactive oxygen species (ROS) and cancer: Role of antioxidative nutraceuticals. *Cancer letters* 2017;**387**: 95-105.
17. Laurent A, Nicco C, Chereau C, Goulvestre C, Alexandre J, Alves A, Levy E, Goldwasser F, Panis Y, Soubrane O, Weill B, Batteux F. Controlling tumor growth by modulating endogenous production of reactive oxygen species. *Cancer research* 2005;**65**: 948-56.
18. Wang H, Gao Z, Liu X, Agarwal P, Zhao S, Conroy DW, Ji G, Yu J, Jaroniec CP, Liu Z, Lu X, Li X, et al. Targeted production of reactive oxygen species in mitochondria to overcome cancer drug resistance. *Nat Commun* 2018;**9**: 562.
19. Grigoryan H, Edmands W, Lu SS, Yano Y, Regazzoni L, Iavarone AT, Williams ER, Rappaport SM. Adductomics Pipeline for Untargeted Analysis of Modifications to Cys34 of Human Serum Albumin. *Analytical chemistry* 2016;**88**: 10504-12.
20. Rappaport SM, Li H, Grigoryan H, Funk WE, Williams ER. Adductomics: characterizing exposures to reactive electrophiles. *Toxicology letters* 2012;**213**: 83-90.
21. Palli D, Berrino F, Vineis P, Tumino R, Panico S, Masala G, Saieva C, Salvini S, Ceroti M, Pala V, Sieri S, Frasca G, et al. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. *Tumori* 2003;**89**: 586-93.
22. Riboli E, Kaaks R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* 1997;**26 Suppl 1**: S6-14.
23. Sandanger TM, Nøst TH, Guida F, Rylander C, Campanella G, Muller DC, van Dongen J, Boomsma DI, Johansson M, Vineis P, Vermeulen R, Lund E, et al. DNA methylation and associated gene expression in blood prior to lung cancer diagnosis in the Norwegian Women and Cancer cohort. *Scientific Reports* 2018: In Press.
24. Leffondre K, Abrahamowicz M, Xiao Y, Siemiatycki J. Modelling smoking history using a comprehensive smoking index: application to lung cancer. *Stat Med* 2006;**25**: 4132-46.
25. Liu S, Grigoryan H, Edmands WMB, Dagnino S, Sinharay R, Cullinan P, Collins P, Chung KF, Barratt B, Kelly FJ, Vineis P, Rappaport SM. Cys34 Adductomes Differ between Patients with Chronic Lung or Heart Disease and Healthy Controls in Central London. *Environ Sci Technol* 2018;**52**: 2307-13.
26. Lu SS, Grigoryan H, Edmands WM, Hu W, Iavarone AT, Hubbard A, Rothman N, Vermeulen R, Lan Q, Rappaport SM. Profiling the Serum Albumin Cys34 Adductome of Solid Fuel Users in Xuanwei and Fuyuan, China. *Environmental science & technology* 2017;**51**: 46-57.
27. Campanella G, Gunter MJ, Polidoro S, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Fiorito G, Guarrera S, Iacoviello L, Bergdahl IA, et al. Epigenome-wide association study of adiposity and future risk of obesity-related diseases. *Int J Obes* 2018.
28. Chadeau-Hyam M, Vermeulen RC, Hebels DG, Castagne R, Campanella G, Portengen L, Kelly RS, Bergdahl IA, Melin B, Hallmans G, Palli D, Krogh V, et al. Prediagnostic transcriptomic markers of Chronic lymphocytic leukemia reveal perturbations 10 years before diagnosis. *Ann Oncol* 2014;**25**: 1065-72.
29. McHale CM, Zhang LP, Lan Q, Vermeulen R, Li GL, Hubbard AE, Porter KE, Thomas R, Portier CJ, Shen M, Rappaport SM, Yin SN, et al. Global Gene Expression Profiling of a Population Exposed to a Range of Benzene Levels. *Environ Health Perspect* 2011;**119**: 628-34.

30. Castagne R, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Kleinjans J, de Kok T, Kyrtopoulos SA, et al. Biological marks of early-life socioeconomic experience is detected in the adult inflammatory transcriptome. *Sci Rep* 2016;**6**: 38705.
31. Castagne R, Delpierre C, Kelly-Irving M, Campanella G, Guida F, Krogh V, Palli D, Panico S, Sacerdote C, Tumino R, Kyrtopoulos S, Hosnijeh FS, et al. A life course approach to explore the biological embedding of socioeconomic position and social mobility through circulating inflammatory markers. *Sci Rep* 2016;**6**: 25170.
32. Chadeau-Hyam M, Ebbels TM, Brown IJ, Chan Q, Stamler J, Huang CC, Daviglus ML, Ueshima H, Zhao L, Holmes E, Nicholson JK, Elliott P, et al. Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification. *Journal of proteome research* 2010;**9**: 4620-7.
33. Castagne R, Boulange CL, Karaman I, Campanella G, Santos Ferreira DL, Kaluarachchi MR, Lehne B, Moayyeri A, Lewis MR, Spagou K, Dona AC, Evangelos V, et al. Improving Visualization and Interpretation of Metabolome-Wide Association Studies: An Application in a Population-Based Cohort Using Untargeted 1H NMR Metabolic Profiling. *J Proteome Res* 2017;**16**: 3623-33.
34. Vermeulen R, Saberi Hosnijeh F, Bodinier B, Portengen L, Liquet B, Garrido-Manriquez J, Lokhorst H, Bergdahl IA, Kyrtopoulos SA, Johansson AS, Georgiadis P, Melin B, et al. Pre-diagnostic blood immune markers, incidence and progression of B-cell lymphoma and multiple myeloma: Univariate and functionally informed multivariate analyses. *Int J Cancer* 2018.
35. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan W, Xu T, Elks CE, Aslibekyan S, Moreno-Macias H, Smith JA, et al. Epigenetic Signatures of Cigarette Smoking. *Circulation Cardiovascular genetics* 2016;**9**: 436-47.
36. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, Krogh V, Tumino R, Sacerdote C, Panico S, Severi G, Kyrtopoulos SA, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics* 2015;**24**: 2349-59.
37. Lee PN, Forey BA, Coombs KJ. Systematic review with meta-analysis of the epidemiological evidence in the 1900s relating smoking to lung cancer. *BMC cancer* 2012;**12**: 385-.
38. Kawanishi S, Hiraku Y, Pinlaor S, Ma N. Oxidative and nitrative DNA damage in animals and patients with inflammatory diseases in relation to inflammation-related carcinogenesis. *Biological chemistry* 2006;**387**: 365-72.
39. Toyokuni S, Okamoto K, Yodoi J, Hiai H. Persistent oxidative stress in cancer. *FEBS letters* 1995;**358**: 1-3.
40. Balansky R, Ganchev G, Iltcheva M, Steele VE, De Flora S. Transplacental antioxidants inhibit lung tumors in mice exposed to cigarette smoke after birth: a novel preventative strategy? *Current cancer drug targets* 2012;**12**: 164-9.
41. van Zandwijk N, Dalesio O, Pastorino U, de Vries N, van Tinteren H. EUROSCAN, a randomized trial of vitamin A and N-acetylcysteine in patients with head and neck cancer or lung cancer. For the European Organization for Research and Treatment of Cancer Head and Neck and Lung Cancer Cooperative Groups. *Journal of the National Cancer Institute* 2000;**92**: 977-86.
42. Van Schooten FJ, Besaratinia A, De Flora S, D'Agostini F, Izzotti A, Camoirano A, Balm AJ, Dallinga JW, Bast A, Haenen GR, Van't Veer L, Baas P, et al. Effects of oral administration of N-acetyl-L-cysteine: a multi-biomarker study in smokers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for*

Cancer Research, cosponsored by the American Society of Preventive Oncology 2002;**11**: 167-75.

For Peer Review

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Figure 1: HRMS data acquired for the adduct of N-acetylcysteine. (A) Isotope distribution for the putative N-acetylcysteine adduct in a plasma sample, (B) MS/MS spectrum of the precursor ion at $m/z = 865.43 (+3)$ acquired in the ion trap. The $\Delta m/z$ used to identify the adducted mass is calculated from the shift in the y^{2+} series.

Figure 2: Manhattan plot summarising the results of the association study relating each of the 42 T3 adducts level in relation to lung cancer disease status, using a univariate linear mixed model approach. Each adduct is represented by its $-\log_{10}$ p-value testing the null hypothesis of no difference in adduct level in cases and controls (Y axis) for each assayed T3 adduct, in function of their m/z ratio (X axis). Results are presented for the model adjusting for technical confounders, age, gender and BMI (Model 1, bullets), and for the model additionally adjusting for smoking status (Model 2, triangle). Adducts with p-values above 0.05 are coloured in grey, and those with p-values below 0.05 in dark red. The blue horizontal dashed line represents the nominal 0.05 significance level and the red and green dashed line represent the permutation-based per-test significance level (α') required to control the family wise error rate (FWER) at a 0.1 level for Model 1 and Model 2, respectively. The plain blue line represents the m/z density assayed in our study along the spectrum.

Figure 3: Distribution of the log-transformed level of the 6 T3 adducts associated with Lung Cancer status. Results are presented for healthy controls (light grey) and lung cancer controls (dark grey) separately.

Figure 4: Bi-partite network representation of the pairwise associations between each of the six lung cancer related T3 adducts and the methylation levels at the 2,670 CpG sites found associated to smoking in Joehanes R *et al*, 2016 and assayed in our study population.

Pairwise association was evaluated using a linear mixed model setting the log-transformed adducts levels as outcome and the methylation M-values as predictor. Results are adjusted for technical confounders, age, gender, and BMI. In order to correct for multiple testing, we only represent edges if the p-value for the regression coefficient linking the adduct level and the DNA methylation M-value is below $0.05/110$, where 110 is the number of principal components required to explain more than 90% of the variance of the original (2,670 dimensional) DNA methylation matrix. For clarity, we only represent CpG sites that are found associated to at least one T3 adduct. The CpG sites are coloured according to the p-value measuring their association with smoking status in our study population, and edges are coloured according the direction of the adduct-CpG site association: red for direct associations ($\beta > 0$), and blue for inverse associations ($\beta < 0$).

Figure 5: Receiver Operating Characteristic (ROC) curves from logistic models regressing the disease risk against the adduct levels of NAC (LC34) (left, figures A and C), and Cys-Gly-(H₂O) (LC33, right, figures B and D). Models were fitted using either lung cancer status, irrespective of the histological subtype (top panel, figures A and B), or restricting the cases to large cell carcinomas for NAC adduct (figure C) and to adenocarcinomas for Cys-Gly-(H₂O) adduct (figure D). In each case we ran three different models: one including smoking exposure as measured by pack-years (in red), one including the level of the adduct of interest (in green), one including both the adduct level and the smoking exposure (in light blue). For the analyses of lung cancer (all subtypes) we also investigated a model including the 76 principal components explaining >90% of the variance of the 2,670 smoking related CpG sites (orange), and the model including both these components and the adduct level (dark blue). We used a subsampling procedure (repeated independently 5,000 times) of 80% of the study population as training set and report the ROC curves and corresponding AUC in

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

the validation set as defined by remaining 20% of the population. The plain ROC curve (and the point estimate of the AUC) correspond to the average performance yielded across the 5,000 subsamples, and the coloured areas (and AUC ranges) reflect the extreme performances yielded across the sub-samples.

For Peer Review

Table 1: Results from the association study of the log-transformed levels of the (N=42) T3 adducts in relation to the two most represented histological subtypes of lung cancer in our study. Results are summarised by the effect size estimate (regression coefficient β) and the corresponding p-value testing the null hypothesis of no association ($\beta=0$). We report results for the model adjusting for technical confounders, age, gender and BMI (Model 1), and for the model additionally adjusting for smoking status (Model 2). For readability, we only report the adducts whose level were found significantly different in cases and controls at a nominal significance level of 0.05 for at least one histological subtype.

	Full sample (96 cases)				Adenocarcinoma (41 cases)				Large-cell carcinoma (20 cases)			
	Model 1		Model 2		Model 1		Model 2		Model 1		Model 2	
	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value	β	p-value
LC1 - Unknown	-0.461	4.10e-02	-0.355	1.44e-01	-0.076	6.81e-01	-0.104	5.87e-01	-1.656	3.26e-04	-1.449	2.41e-03
LC10 - Unknown	-0.652	2.18e-02	-0.579	6.01e-02	-0.522	1.13e-01	-0.479	1.61e-01	-1.33	1.36e-02	-1.172	3.84e-02
LC19 - Benzaldehyde	-0.502	1.22e-01	-0.414	2.40e-01	-0.885	4.74e-02	-0.851	6.59e-02	-0.805	1.65e-01	-0.726	2.37e-01
LC26 - Unknown	-0.368	3.19e-02	-0.329	7.58e-02	-0.27	2.16e-01	-0.288	2.05e-01	-0.709	1.98e-02	-0.62	5.14e-02
LC33 - S-Cys-Gly -(H2O)	-0.489	7.89e-03	-0.51	1.07e-02	-0.587	2.22e-02	-0.633	1.77e-02	-0.505	7.50e-02	-0.48	1.09e-01
LC34 - S-(Nacetyl)Cys	-0.587	4.15e-03	-0.505	2.25e-02	-0.229	2.85e-01	-0.154	4.87e-01	-1.219	3.41e-04	-1.199	8.88e-04
LC35 - S-Cys-Gly	-0.315	2.56e-02	-0.302	4.85e-02	-0.069	6.81e-01	-0.119	4.97e-01	-0.554	1.95e-02	-0.627	1.23e-02

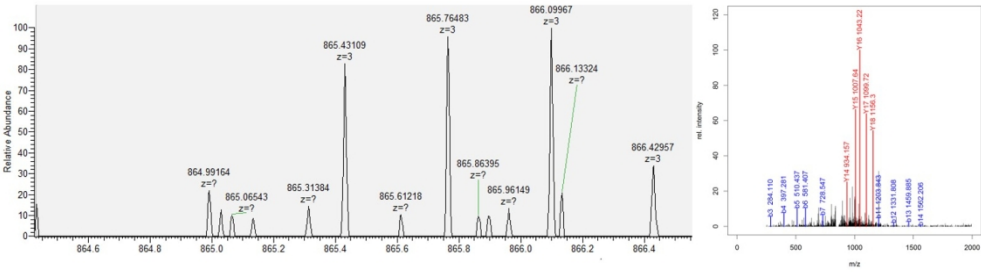


Figure 1: HRMS data acquired for the adduct of N-acetylcysteine. (A) Isotope distribution for the putative N-acetylcysteine adduct in a plasma sample, (B) MS/MS spectrum of the precursor ion at m/z= 865.43 (+3) acquired in the ion trap. The $\Delta m/z$ used to identify the adducted mass is calculated from the shift in the y^{2+} series

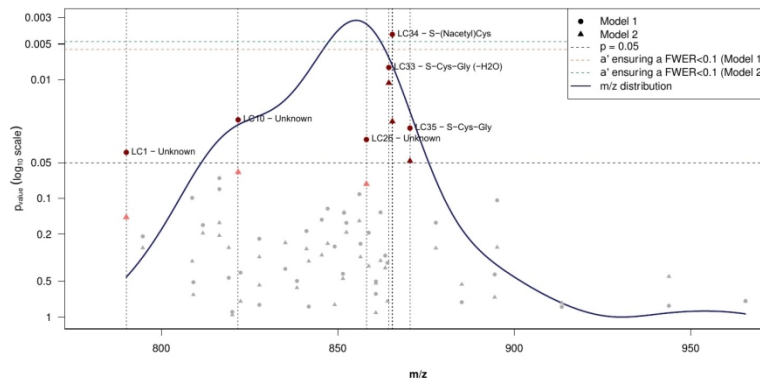


Figure 2: Manhattan plot summarising the results of the association study relating each of the 43 T3 adducts PAR in relation to lung cancer disease status, using a univariate linear mixed model approach. Each adduct is as the $-\log_{10}$ p-value testing the null hypothesis of no difference in adduct level in cases and controls (Y axis) for each assayed T3 adduct, in represented by their m/z ratio (X axis). Results are presented for the model adjusting for technical confounders, age, gender and BMI (Model 1, bullets), and for the model additionally adjusting for smoking status (Model 2, triangle). Adducts with p-values below 0.05 are coloured in grey, and those with p-values <0.05 in dark red. The blue horizontal dashed line represents the nominal 0.05 significance level and the red and green dashed line represent the permutation-based per-test significance level (α') required to control the family wise error rate (FWER) at a 0.1 level for Model 1 and Model 2, respectively. The plain blue line represents the m/z density assayed in our study along the spectrum.

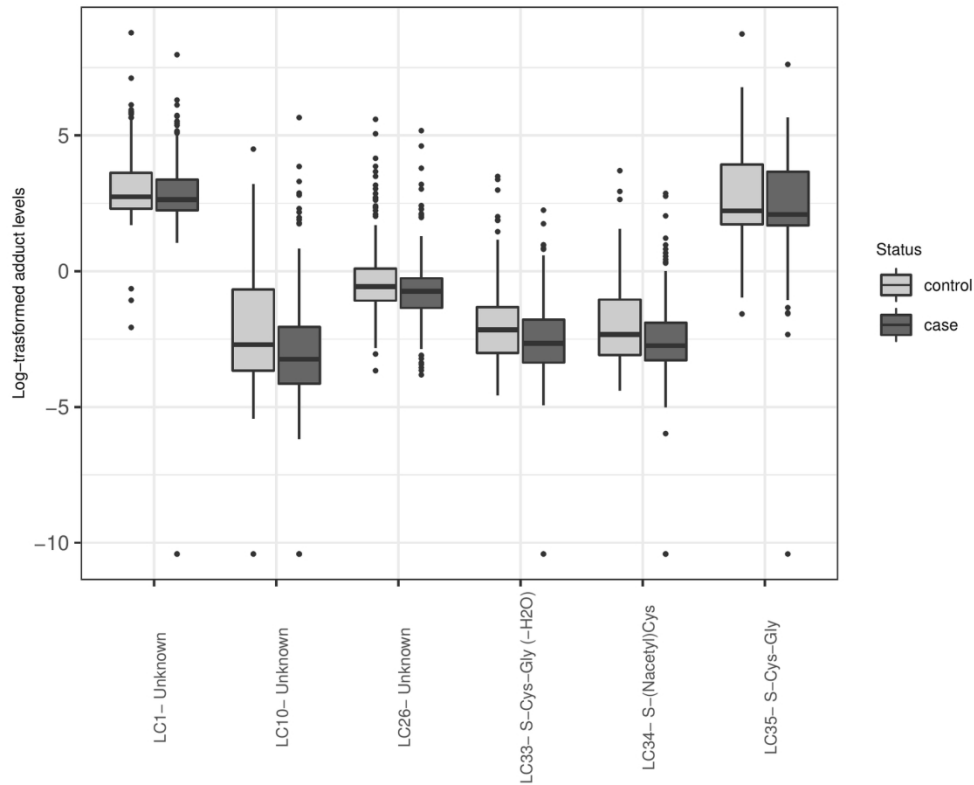
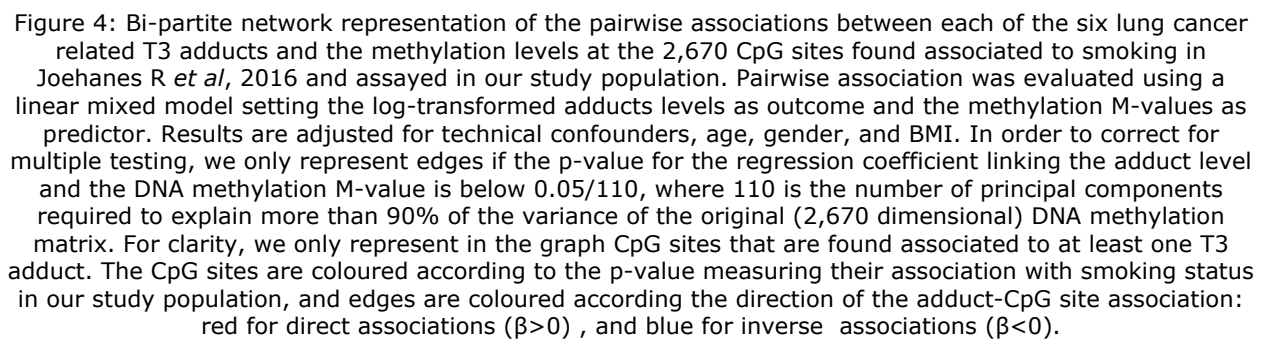


Figure 3: Distribution of the log-transformed level of the 6 T3 adducts associated with Lung Cancer status. Results are presented for healthy controls (light grey) and lung cancer controls (dark grey) separately.

152x127mm (300 x 300 DPI)



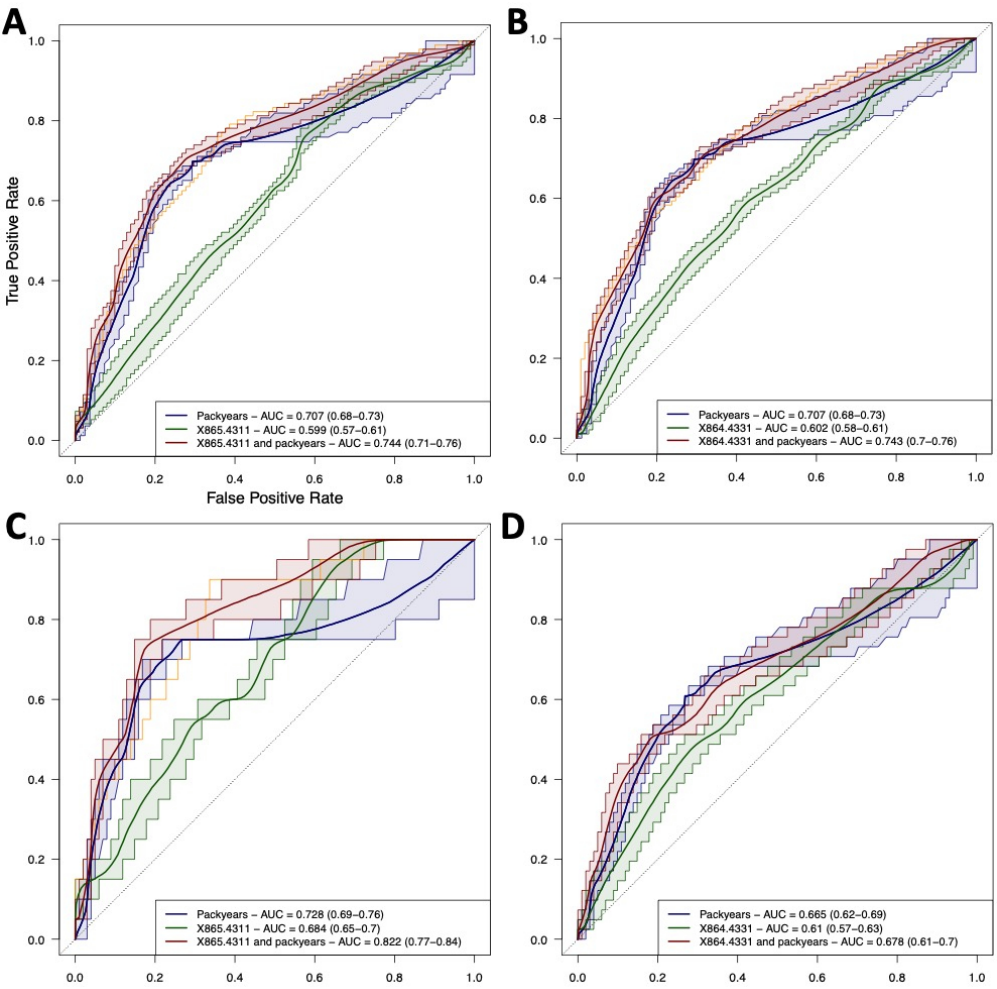


Figure 5: Receiver Operating Characteristic (ROC) curves from logistic models regressing the disease risk against the PAR levels of to X865.4311 (left, figures A and C), and X864.4331 (right, figures B and D). Models were fitted using either lung cancer status, irrespective of the histological subtype (top panel, figures A and B), or restricting the cases to large cell carcinomas for X865.4311 (figure C) and to adenocarcinomas for X865.4331 (figure D). In each case we ran three different models: one including smoking exposure as measured by packyears (in blue), one including the PAR levels of the T3 adduct of interest (in green), one including both the adduct and the smoking exposure (in red). We used a subsampling procedure (repeated independently 5,000 times) of 80% of the study population as training set and report the ROC curves and corresponding AUC in the validation set as defined by remaining 20% of the population. The plain ROC curve (and the point estimate of the AUC) correspond to the average performance yielded across the 5,000 subsamples, and the coloured areas (and AUC ranges) reflect the extreme performances yielded across the subsamples.