Stochastic epigenetic mutations are associated with risk of breast cancer, lung cancer and mature B-cell neoplasms

Amedeo Gagliardi[1,2+] Pierre-Antoine Dugué[3,4,5,+], Therese H Nøst[7+], Melissa C. Southey[3,5,6] ,Daniel D Buchanan[4,8,9], Daniel F Schmidt[4,10], Enes Makalic[4], Allison M Hodge[3,4], Dallas R English[3,4], Nicole W Doo[3,11,12] ,John L Hopper[4], Gianluca Severi[13], Laura Baglietto[13,14],Alessio G Naccarati[1,2], Sonia Tarallo[1,2], Luigia Pace[1], Vittorio Krogh[15], Domenico Palli[16], Salvatore Panico[17], Carlotta Sacerdote[18], Rosario Tumino[19], Eiliv Lund[7], Graham G Giles[3,4,5], Barbara Pardini[1,2], Torkjel M Sandanger[7,*], Roger L Milne[3,4,5*], Paolo Vineis[1,20*], Silvia Polidoro[1,20*], Giovanni Fiorito[20,21*]

[1]Italian Institute for Genomic Medicine (IIGM, former HuGeF), c/o IRCCS Candiolo, SP142, km 3,95 – 10060 Candiolo.
[2]Candiolo Cancer Institute, FPO – IRCCS, Candiolo (IT)
[3]Cancer Epidemiology Division, Cancer Council of Victoria, 615 St Kilda Road, Melbourne, Victoria, 3004, Australia.
[4]Centre for Epidemiology and Biostatistics, School of Population and Global Health, The University of Melbourne, Parkville Victoria 3010, Australia.
[5]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia.
[6]Department of Clinical Pathology, The University of Melbourne, Parkville Victoria 3010, Australia.
[7]Department of Community Medicine, Faculty of Health Sciences, UiT-The Arctic University of Norway, NO-9037 Tromsø, Norway[8]Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Parkville, Victoria, Australia.
[8]Colorectal Oncogenomics Group, Department of Clinic Pathology, The University of Melbourne, Victorian Comprehensive Cancer Centre, Melbourne, Victoria, Australia.
[9]Genomic medicine and Family Cancer Clinic, Royal Melbourne Hospital, Melbourne, Victoria, Australia.
[10]Faculty of Information Technology, Monash University, Victoria, Australia.
[11]Concord Repatriation General Hospital, Sydney Medical School, University of Sydney, NSW 2139, Australia.
[12]Concord Clinical School, University of Sydney, Concord, NSW 2139, Australia
[13]Centre de Recherche en Épidémiologie et Santé des Populations (CESP, Inserm U1018), Université Paris-Saclay, UPS, USQ, Gustave Roussy, Villejuif, France.
[14]Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy.
[15]Fondazione IRCCS - Istituto Nazionale dei Tumori, Via Venezian 1, 20133, Milan, Italy.
[16]Institute for Cancer Research, Prevention and Clinical Network - ISPRO, Villa delle Rose, Via Cosimo il Vecchio, 2 -50139 Florence, Italy.
[17]Department of Clinical Medicine and Surgery, University of Naples Federico II, Corso Umberto I, 40, 80138, Naples, Italy.
[18]Piedmont Reference Centre for Epidemiology and Cancer Prevention (CPO Piemonte), Via Santena 7, 10126, Turin, Italy.
[19]Cancer Registry and Histopathology Department, 'Civic - M. P. Arezzo' Hospital, ASP Ragusa, Piazza Igea, 1, 97100, Ragusa, Italy.

[20]MRC-PHE Centre for Environment and Health, Imperial College London, St. Mary's Campus Paddington, W2 1PG London, United Kingdom.
[21] Laboratory of Biostatistics, Department of Biomedical Sciences, University of Sassari
[+,*] equal contribution

Keywords: DNA methylation, Stochastic epigenetic mutations, Cancer epigenetics

Financial support:

This research was supported by the 'Lifepath' grant awarded to Imperial College, London, and the Italian Institute for Genomic Medicine (IIGM) in Turin, Italy and AIRC grant (Progetto IG 2013 N.14410) to Carlotta Sacerdote for part of the DNA methylation experiments. The Melbourne Collaborative Cohort Study cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS component of the work was funded by the Australian National Health and Medical Research Council, including grants 1106016, 1011618, 1026892, 1027505, 1050198, 1087683, 1088405, 1043616, 209057, 396414 and 1074383. Cases and their vital status were ascertained through the Victorian Cancer Registry and the Australian Institute of Health and Welfare, including the National Death Index and the Australian Cancer Database. The NOWAC component of the work was supported by the European Research Council (ERC) Advanced Researcher Grant, 2008: Transcriptomics in cancer research (TICE).

Corresponding author:

Amedeo Gagliardi, gagliardi.borsisti@iigm.it

Conflict of interest

The authors declare no conflict of interest.

Manuscript additional info

Words count: 3225

Commented [GF1]: Ame aggiornare alla fine

Figures count: 2

Tables count: 3

Abbreviations

EPIC: The European Prospective Investigation into Cancer and nutrition,

MCCS: The Melbourne Collaborative Cohorts Study,

NOWAC: The NOrwegian Women And Cancer study,

AA: age acceleration,

SEM: stochastic epigenetic mutation,

DNAm: DNA methylation,

UCC: urothelial cell carcinoma,

MBCN: mature B-cell neoplasm,

WBC: white blood cell,

log(SEM): logarithm of the total number of SEMs,

BMI: body mass index,

OR: odds ratio,

CI: confidence interval,

PRC2: Polycomb-Repressive-Complex-2,

ChIP-Seq: Chromatin ImmunoPrecipitation Sequencing,

CNV: copy number variant,

TFBS: transcription factor binding site,

IEAA: intrinsic epigenetic age acceleration,

IQR: interquartile range,

TTD: time to disease,

sd: standard deviation.

Abstract

Background: Age-related epigenetic dysregulations were associated with several diseases, including cancer. The individual number of stochastic epigenetic mutations (SEMs) has been suggested as a biomarker of life-course accumulation of exposure-related DNA damage; however, the predictive role of SEMs in cancer has seldom been investigated.

Methods: A SEM, at a given CpG site, was defined as an extreme outlier of blood DNA methylation value distribution across individuals. We investigated the association of the total number of SEMs with the risk of eight cancers in 4,497 case-control pairs nested in three prospective cohorts. Further, we investigated whether SEMs were randomly distributed across the genome or enriched in functional genomic regions.

Results: In the three-study meta-analysis the estimated odds ratios (ORs) per one-unit increase in log(SEM) from logistic regression models adjusted for age and cancer risk factors were 1.25; 95% CI 1.11-1.41 for breast cancer, and 1.23; 95% CI 1.07-1.42 for lung cancer. In MCCS, the OR for mature B-cell neoplasm was 1.46; 95% CI 1.25-1.71. Enrichment analyses indicated that SEMs more likely occur in silenced genomic regions and in transcription factor binding sites regulated by EZH2 and SUZ12 ($p<0.0001$ and $p=0.0005$ respectively): two components of the Polycomb-Repressive-Complex-2 (PCR2). Finally, using longitudinal DNA methylation data, we showed that PCR2-specific SEMs are generally more stable in time compared with SEMs occurring in the whole-genome.

Conclusions: The number of SEMs is associated with a higher risk of different cancers in pre-diagnostic blood samples. Enrichment analyses indicate key enzymatic pathways possibly involved in carcinogenesis mechanisms.

Impact: We provide the first evidence of the prospective association between epimutations and a higher risk of different cancers. We hypothesized a possible mechanism of carcinogenesis involving PCR2 complex proteins worthy of further investigation.

Introduction

The concept of 'life-course accumulation of exposures' and related damage has been proposed to explain the decline of physiological functioning and the consequent increased disease morbidity and mortality during aging(1). The accumulation of environmental, socioeconomic and behavioural exposures may cause long-term damage, which may be amplified by a decreased ability to repair damage as the body ages(1). Age is, in fact, an important risk factor for most diseases, including cancer, and the incidence of most cancers increases exponentially with age(2).

Basic research, combined with the increasing capacity of large-scale technologies including 'omics' measurements, has led to the formulation of exposure-driven models of carcinogenesis(3), in which functional changes in gene regulation and genomic mutations reflect the life-course accumulation of exposure-related DNA damage. It has long been postulated that the accumulation over time of somatic mutations in specific genes may lead to cancer development, but recent studies demonstrated that this molecular mechanism alone is not sufficient(4,5).

Epigenetic landscapes, in particular, change considerably across the individual lifespan, suggesting that epigenetic variability is a fundamental component of the aging process(4,6), constituting a link between genetic and environmental factors via the regulation of gene transcription processes. DNA methylation (DNAm) is the most studied epigenetic mechanism, and changes in DNA methylation over time are thought to play a role in several age-related diseases, including cancer(6),(7).

Two mechanisms contribute to age-related DNA methylation changes: the 'epigenetic drift'(6) and the 'epigenetic clock'(8). Although both are related to aging, the 'epigenetic clock' refers to specific CpG sites at which DNA methylation levels steadily increase or decrease with age and thus can be used to predict chronological age with high accuracy(8). The concept of epigenetic age acceleration has been introduced as the difference between predicted DNA methylation age and the chronological age(8,9). Epigenetic age acceleration may be a good biomarker of biological aging as it has been associated with longevity(10-13), several pathological conditions(14,15), and

non-communicable disease risk factors like obesity(16), poor physical activity(17), and low socioeconomic status(18). Previous work found a consistent association between measures of epigenetic aging and increased cancer risk and shorter cancer survival (11). Recent literature discerns Horvath (8) and Hannum (9) 'first-generation clocks' from DNAmPhenoAge (19) and DNAmGrimAge (20), called the 'next-generation clocks', the latest being trained not only on age instead, on a complex set of biomarkers which in turn are associated with individual health status and mortality. Early findings seem to indicate that the next-generation clocks may be capturing important aspects of accelerated biological aging. In a recent critique of the epigenetic clocks, Dugue et al. cautioned that early studies generally report stronger associations than later studies and are more likely to be affected by publication bias (21).

In contrast, 'epigenetic drift' is a mechanism that involves the whole-genome, suggesting a global dysregulation of DNA methylation patterns with age(22). Two critical aspects of the epigenetic drift are genomic instability and chromatin deterioration during aging, which lead to an accumulation of epigenetic mutations (also known as 'epimutations', i.e. changes in gene activity not involving DNA mutations but rather gain or loss of DNA methyl groups, which are conserved in cells during mitosis(23)). A higher number of stochastic epigenetic mutations (SEMs) across the genome has been associated with risk factors such as cigarette smoking, alcohol intake(23) and exposure to toxicants(24). We recently reported several associations between lifestyle-related variables and the number of SEMs (25). Moreover, more SEMs may be associated with skewed X chromosome inactivation in women and with hepatocellular carcinoma tumour stage(26) suggesting a possible role of SEMs in other age-related diseases.

In this study, we investigated the associations between the number of SEMs across the genome and the risk of eight malignancies (breast, colorectal, lung, gastric, prostate, and kidney cancer, as well as urothelial cell carcinoma (UCC), and mature B-cell neoplasms (MBCN)) in 4,497 case-controls pairs, matched on age and other relevant variables, nested within three large cohorts from Italy (the Italian part of the European Prospective Investigation into Cancer and Nutrition

Study (EPIC)), Australia (the Melbourne Collaborative Cohort Study (MCCS)), and Norway (the Norwegian Women and Cancer Study (NOWAC)). This is the first prospective study to assess the association between the number of SEMs and cancer risk in DNA derived from blood samples.

Before this study, only Teschendorff et al. (27) investigated such relationship in cancer cells. We also investigated the biomolecular mechanisms linking aging, DNA methylation patterns, and the risk of different cancers analyzing the genome-wide distribution of epimutations, to identify functional genomic regions enriched in SEMs, and to describe the biomolecular mechanism of carcinogenesis possibly.

Methods

Study sample

Details of participant recruitment and relevant covariate acquisition are reported in the supplementary text. Briefly, EPIC Italy, MCCS and NOWAC are prospective cohort studies with demographic and lifestyle variables and blood samples collected from participants at recruitment. For each cohort, subsets of blood samples were previously selected for DNA methylation analyses, using nested case-control study designs, using the incidence density sampling method for case-control matching (11,28-30). In EPIC Italy, three sub-study samples were case-control studies on breast, lung and colorectal cancer (556 cases and controls, 45% breast cancer, 30% lung cancer, 25% colorectal cancer). The median time to disease (TTD) were: 7.01 years (interquartile range (IQR) = 7.09), 7.44 years (IQR = 5.65), and 6.28 years (IQR = 5.04) for breast, lung, and colorectal cancer studies respectively. Case-control pairs were matched by age (±2.5 years), sex, season of blood collection, centre of recruitment, and length of follow-up. The average age difference in absolute value between cases and matched controls was 0.25 (standard deviation 0.26). In NOWAC, two sub-study samples were case-control studies on breast and lung cancer (316 cases and controls, 59% breast cancer, 41% lung cancer). For each case, one control with adequate blood samples was selected matched on time since blood sampling and year of birth (that is cases and matched controls had the same age at recruitment) in order to control for effects of storage time and

196 age. The median TTD were: 2.10 years (IQR = 2.14) and 4.10 years (IQR = 3.21) for breast and
197 lung cancer study, respectively. The average age difference in absolute value between cases and
198 matched controls was 0.13 (standard deviation 0.33). Finally, in MCCS eight sub-studies were on
199 breast, lung, colorectal, gastric, kidney and prostate cancer, UCC and MBCN (3,625 cases and
200 controls, 11% breast cancer, 9% lung cancer, 23% colorectal cancer, 5% gastric cancer, 4% kidney
201 cancer, 24% prostate cancer, 12% UCC, 12% MBCN). For each nested case-control study, controls
202 were individually matched to incident cases on age (±2.5 years), sex, country of birth, blood DNA
203 source and collection period. The average case-control age difference in absolute value was XX (sd
204 = XX). The median TTD were 7.7 years (IQR = 6.07), 9.3 years (IQR = 7.9), 11.4 years (IQR =
205 10.3), 11.2 years (IQR = 8.5), 10.1 years (IQR = 7.5), 10.5 years (IQR = 8.1), 10.5 years (IQR =
206 7.9), 6.3 years (IQR = 6.8) for breast, colorectal, gastric, kidney, lung, MBCN, prostate and UCC
207 study respectively.
208 A total of 4,497 case-control matched pairs were analyzed (**Table 1**).
209 This study was conducted following the principles of the Declaration of Helsinki and its
210 subsequent revisions, and all study participants signed informed consent. EPIC was reviewed and
211 approved by the HuGeF (currently IIGM) Ethics Committee. The MCCS protocol was approved by
212 the Cancer Council Victoria's Human Research Ethics Committee. NOWAC was approved by the
213 Regional Committee for Medical and Health Research Ethics in North Norway.
214 *DNA methylation analyses*
215 Whole-genome DNA methylation was quantified using the Illumina Infinium
216 HumanMethylation450 BeadChip. Detailed methods and data pre-processing procedures can be
217 found in the supplementary text. To account for the possible bias introduced by the inter-individual
218 variability in the proportion of white blood cells (WBC) in peripheral blood, we estimated the
219 percentage of WBC fractions according to the Houseman algorithm(31), which performs inference
220 using a quadratic programming technique known as linear constrained projection, where non-
221 negativity and normalization constraints on cellular proportions are imposed during inference(32).

Commented [GF2]: Pierre, please double-check.

Commented [GF3]: Pierre, please fill.

We excluded from the analysis bimodal and trimodal CpGs using the function *findpeaks* in the R package *pracma*, thus focusing on rare, stochastic events. Missing methylation values were imputed using the k-nearest neighbours algorithm using the R function *impute.knn(33).*

Statistical analyses

Identification of stochastic epigenetic mutations.

We computed the total number of SEMs as the sum of extreme DNA methylation values (outliers) per individual. This approach, based on a modified version of the procedure described by Gentilini et al.,(34) take into account differential WBC proportions among individuals. Specifically, for each CpG, we computed the residuals from the regression of DNA methylation beta values on estimated WBC fractions and then, considering the distribution of DNA methylation beta values across all samples, we computed the interquartile range (IQR) – the difference between the 3rd quartile (Q3) and the 1st quartile (Q1) for the residuals - and defined a SEM as a methylation value lower than Q1-(3×IQR) or greater than Q3+(3×IQR). Finally, for each individual, we computed the total number of SEMs across the assay. The described procedure leads to an estimation of the total number of SEMs per individual independent on individual differential WBC proportion by definition. In **Figure S1,** we show the Spearman correlation coefficients of the total number of SEMs with estimated WBC percentages. Since the number of SEMs increased exponentially with age, we used a logarithmic transformation of the total number of SEMs (referred to hereafter as log(SEM)) for all association analyses.

Computation of epigenetic clock measures.

We computed two measures of epigenetic age acceleration (AA) based on Horvath DNAmAge(8) and DNAmGrimAge(20) according to the algorithm described by Horvath and colleagues. Briefly, DNAmAge was calculated as a weighted average of 353 age-related CpGs (Horvath DNA methylation age). Weights are defined using a penalized regression model (Elastic-net regularisation) (8). Age acceleration (AA) was defined as the difference between epigenetic and chronological age. Since AA may be correlated with chronological age and WBC proportions, we

also computed the 'intrinsic epigenetic age acceleration' (IEAA), defined as the residuals from the linear regression of AA on chronological age and WBC percentages (13). Positive values of IEAA (which by definition is independent of age and WBC) indicate accelerated aging and vice versa. The DNAmGrimAge also known as the 'next-generation clock', is a composite biomarker based on DNAm surrogate measures of seven plasma proteins associated with overall mortality in addition to DNAm surrogate of smoking pack-years, trained to be strongly predictive of overall mortality. The methods for enrichment analyses of the identified epimutated CpGs are described in the Supplementary Material.

Association of SEMs with cancer risk.

We investigated the association between SEMs and the risk of eight types of cancer separately using log(SEM) as the predictor and case-control status as the outcome. Odds ratios (ORs) and confidence intervals (CIs) were calculated using conditional logistic regression models for a one-unit increase in log(SEM). For each cancer and each cohort, we ran four regression models: Model 1 included age, sex, and study-specific covariates (centre of recruitment in EPIC, ethnicity and tissue type in MCCS); Model 2 included additional adjustment for cancer risk factors: smoking, body mass index (BMI), physical activity, alcohol intake, dietary quality and education (as a proxy for socioeconomic status); Model 3 included additional adjustment for Horvath epigenetic AA; finally, Model 4 included additional adjustment for DNAmGrimAge epigenetic age acceleration. All covariates were treated as categorical variables with three categories to harmonize sources of information across the three studies (see Supplementary Material for more details on harmonization of covariates).

Commented [GF4]: Pierre, please double-check

For associations with breast, lung and colorectal cancer, which were investigated in more than one study, the overall OR estimates for the association between log(SEM) and cancer risk were calculated using random-effect maximum likelihood (REML)(35) meta-analysis using the R package *metafor(36)*. Heterogeneity in the associations among studies was evaluated using the I^2 statistic. Further sensitivity analyses were performed stratifying case-control pairs based on the case

time between blood collection and cancer diagnosis (time to disease (TTD)); ORs and confidence intervals were computed on subsample having TTD > 10 years, TTD between 5 and 10 years, and TTD <= 5 years. Cochran-Armitage test for trend was used to evaluate ORs increase with decreasing TTD.

SEMs stability over time.

To evaluate the stability of SEMs over time, we analyzed DNAm data from the Italian part of the Personal Exposure Monitoring (PEM-Turin) study, which in turn is part of the EXPOsOMICS project(37). The PEM-Turin study included 42 healthy volunteers, whose whole-genome DNAm was measured twice in 2015 as part of a study aimed at investigating the effect of air pollution exposure on 'omic' biomarkers(38). Thirty-three out of 42 volunteers were already enrolled in the EPIC Italy study in the '90s and are part of this study sample as healthy controls. That is, we were able to compare epimutation patterns at the time of recruitment in EPIC Italy, with epimutation patterns around 19 years later (mean = 18.75 years, range = 16.45 - 20.26 years) using longitudinal data.

SEMs in cancer tissues.

We evaluated the consistency of epimutation patterns identified in blood pre-diagnostic samples with tissue-specific (both normal and cancerous) epimutation profiles. Data from The Cancer Genome Atlas (TCGA) project were downloaded from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov); specifically, we investigated epimutation profiles on tumoral-normal adjacent tissue pairs from 32 lung cancer patients (TCGA-LUAD project), 91 breast cancer patients (TCGA-BRCA project), and 45 colorectal cancer patients (TCGA-COAD and TCGA-READ project).

Data availability

The data generated and/or analyzed in the current study could be accessed upon reasonable request to the originating cohorts. Access will be conditional to adherence to local ethical and

security policy. R codes used for the analyses presented in the paper are available upon request. EPIC DNAm partial data can be accessed through GEO accession number GSE51057.

Results

Association of cancer risk factors with SEMs

Analyzing the number of SEMs in the 3 cohorts, we observed an exponential increase in the number of SEMs with age both in the whole study sample (**Figure 1**; Pearson R=0.17, $p=5x10^{-9}$; R=0.04, $p=6x10^{-5}$; R=0.23, $p=2x10^{-9}$ in EPIC, MCCS and NOWAC, respectively) and in controls only (**Figure 1**; Pearson R=0.15, $p=2x10^{-5}$; R=0.04, p=0.01; R=0.23, $p=1x10^{-8}$ in EPIC, MCCS and NOWAC, respectively). In **Table 2** are reported the cross-sectional associations of cancer risk factors with log(SEM) in both the whole study sample and in controls only. In EPIC Italy, log(SEM) was associated with smoking status, BMI and education in the whole study sample, and with BMI only in controls sample. In MCCS log(SEM) was associated with BMI, physical activity and education in the whole sample and with … in controls only. No association was observed in NOWAC. In both MCCS and EPIC, log(SEM) was greater in obese individuals; in EPIC, log(SEM) was greater in current smokers and the low education group. In the MCCS, log(SEM) was lower in the low education group and among individuals with low physical activity.

Commented [GF5]: Pierre and Therese, please fill Table 2.

Association of SEMs with the risk of cancers

In the regression Model 2, adjusting for major cancer risk factors, the presence of more SEMs was associated with an increased risk of breast cancer (meta-analysis: OR per one-unit increase in log(SEM)=1.25; 95% CI 1.11-1.41; p=0.0003; I^2=0%; **Figure 2a**), and lung cancer (meta-analysis: OR=1.23; 95% CI 1.07-1.42; p=0.004; I^2=0%; **Figure 2b**). No association was found in the meta-analysis of colorectal cancer in EPIC and MCCS (OR=1.02; 95% CI 0.91-1.14; p=0.74; I^2=0%; **Figure 2c**). In MCCS only, log(SEM) was associated with MBCN (OR=1.43; 95% CI 1.22-1.67; $p=5x10^{-06}$, **Table 3**). ORs greater than one per log(SEM) were also observed for kidney and prostate cancers, although the associations were not statistically significant (**Table 3**).

Interestingly, the ORs from Model 1 did not deviate significantly from those estimated in Model 2 (**Table 3**), and evidence of association with risk of breast and lung cancers and MBCN was observed, after adjustment for smoking, BMI, alcohol intake, diet and education as covariates in the

logistic regression models, suggesting limited confounding by these variables. Similarly, additional adjustments for the epigenetic clock measures in Model 3 and Model 4 did not change the estimated ORs significantly (Table 3). In the analysis stratified by TTD, we found a significant increase in ORs as the TDD decrease for breast, colorectal (p for trend < 0.001), MBCN, and prostate cancer (p for trend < 0.05, **Figure S2**).

Association of number of SEMs with epigenetic clocks.

As shown in **Figures S3 and S4**, the number of SEMs was positively correlated with Horvath DNAmAge epigenetic clock in all three studies (R = 0.25, p < 0.0001; R = 0.03, p = 0.001; R = 0.20, p = 0.04 in EPIC, MCCS and NOWAC, respectively), and with GrimDNAmAge (R = 0.25, p=0.0005; R = 0.07, p<0.0001; R =0.24, p=0.04 in EPIC, MCCS and NOWAC, respectively). Consistent results were obtained from the analyses of control sample only.

Enrichment analyses

We investigated enrichment of SEMs in specific genomic regions based on the Illumina annotation about CpG site location. We found enrichment of epimutations in genomic regions characterized by open chromatin states, CpG islands and shores (p=0.02, p=0.05 and p=0.0003 respectively, **Table S1**). Considering the functional categories defined by the ENCODE project with Chromatin Immuno Precipitation Sequencing (ChIP-Seq) experiments on human embryonic stem cells (hESC), we found enrichment of SEMs in 'inactive/poised promoters' (p<0.0001), 'heterochromatin/low signal/CNV' (p<0.0001), and 'Polycomb-repressed' regions (p=0.02) (**Table S2**). Furthermore, considering transcription factor binding sites (TFBSs) in hESC from ENCODE project, we also found significant an enrichment of SEMs in TFBSs targeted by two members of the Polycomb-Repressive-Complex-2 (PRC2): EZH2 and SUZ12 (p<0.0001 and p=0.0005, respectively, **Table S3**) and by the transcriptional corepressor ctBP2 (p=0.001, **Table S3**).

Association of EZH2-specific SEMs with the risk of cancer

Given the enrichment analysis results, we further investigated SEMs in EZH2 targets (in which the evidence for enrichment was the strongest). The number of SEMs in regions targeted by

EZH2 was strongly correlated with the total number of SEMs across all the genome (Pearson R >0.80, **Figure S5**). We repeated the tests for the associations with cancer, considering the EZH2-specific SEMs and obtained results consistent with those presented in **Table 3**; EZH2-specific SEMs were strongly associated with breast cancer, lung cancer and MBCN (**Table S4**). Adjustment for batch effects did not substantially influence the association observed ('*Supplementary results',* Supplementary Material). It is worth observing that the majority of the CpG sites targeted by EZH2 were on average hypo-methylated (more than 80% of the CpGs have average DNAm beta value lower than 20%, **Figure S6**); consequently, more than 95% of EZH2-specific SEMs occur as abnormal hyper-methylation of a locus that is hypo-methylated in the overall sample.

SEMs stability over time

In the longitudinal regression model on PEM-Turin dataset, the total number of SEMs per individual significantly increased in time (log(SEM) increase per year = 0.168 ± 0.007; $p < 0.0001$, **Figure S7**. Among the epimutations identified at baseline, the majority were still present at the time of PEM-Turin study (18.75 years later, on average, range = 16.45 - 20.26 years). The average percentage of conserved SEMs was 71% (range 55% - 93%). Based on the results of the enrichment analyses, we focused on EZH2-specific epimutations. The proportion of conserved EZH2-specific epimutations was significantly higher compared with what observed at genome-wide level (mean = 87%; range = 62% - 100%; Chi-Squared test for proportion $p < 0.0001$).

SEMs in tumour compared with normal adjacent tissues

To verify the consistency among the results obtained in pre-diagnostic blood samples with epimutation patterns in cancer tissues, we analyzed data from the TCGA project on lung, breast and colorectal cancers. The differences in log(SEM) between cancer and normal adjacent tissues were 4.11 (95% CI 3.70 – 4.52; paired Student T-test $p < 0.0001$) for lung cancer; 3.29 (95% CI 2.98 – 3.62; $p < 0.0001$) for breast cancer; 3.94 (95% CI 3.54 – 4.33; $p < 0.0001$) for colorectal cancer (**Figure S8 a, b, c**). The observed differences were even higher looking at EZH2-specific SEMs: 5.37 (95% CI 4.77 - 5.94; $p < 0.0001$) for lung cancer; 4.02 (95% CI 3.62 – 4.42; $p < 0.0001$) for

breast cancer; 4.86 (95% CI 4.43 – 5.30; $p < 0.0001$) for colorectal cancer (**Figure S8 d, e, f**). The average proportion of SEMs conserved in tumour from normal-adjacent tissue was 72% (range 54% - 98%); whereas the proportion of conserved EZH2-specific SEMs was significantly higher: 87% (range 61% – 97%, Chi-Squared test for proportion $p < 0.0001$). Finally, enrichment analyses confirmed SEMs more likely occur in silenced genomic regions like inactive and poised promoters, Polycomb repressed regions, and in TFBS of EZH2 and SUZ12.

Discussion

In the present study, we have analyzed DNAm data from blood samples of ~4,500 cancer cases and one-to-one matched controls, nested within three large cohorts: EPIC Italy, MCCS and NOWAC. The main aim of this study was to investigate the association of the total number of SEMs with cancers using a prospective study design. In addition, we investigated SEMs stability over time and genomic regions in which SEMs more likely appear.

SEMs increasing with aging and stability over time

The number of estimated SEMs per sample varied by cohort; however, we observed an exponential increase of SEMs with age in all cohorts (**Figure 1**) confirming the results of previous studies(34,39). Differences in the number of SEMs between studies were mainly driven by batch effect, different normalization and DNAm data pre-processing procedure, and different study sample size which affect CpGs DNAm values distribution, making the comparison of SEMs between different batches challenging. Consequently, the magnitude of the association of logSEM with age (**Figure 1**) and epigenetic clocks (**Figures S3 and S4**) varied by cohort also. Nevertheless, in this study, we aimed to investigate the association of SEMs with cancer, and our study design using matched case-control pairs analyzed in the same batch overcome batch effect issues. The ORs for breast, lung, and colorectal cancer (investigated in more than one cohort) were estimated trough a random effect meta-analysis.

The results observed in our cross-sectional study and reported in the literature about the exponential increase of SEMs with age were further confirmed using longitudinal data, available for

405 a subset of the EPIC Italy study included in the EXPOsOMICS study also. We observed high
406 interindividual variability of the total number the grow rate of SEMs among individual of the same
407 age (**Figure S7**), strengthening our hypothesis of SEMs as candidate biomarkers of accumulation of
408 exposure-related DNA damage during aging, and as a possible biomarker for age-related diseases.
409 Accordingly, in this study sample we observed cross-sectional association of SEMs with lifestyle-
410 related factors like smoking and obesity, and in our previous study with higher sample size with
411 alcohol intake, and socioeconomic status(25). Also, logSEM positively correlates with the widely
412 studied biological aging measures based on the epigenetic clock developed by Horvath and
413 colleagues (**Figure S3 and S4**). The association between the two age-related biomarkers is not
414 driven by their association with chronological age, because the Intrinsic Epigenetic Age
415 Acceleration (IEAA) is independent of chronological age by definition (13).
416 We were not able to investigate whether changes in lifestyle may slow down aging-related
417 SEMs rise using longitudinal data due to the lack of statistical power. A recent study analyzing
418 longitudinal data on SEMs in twins concluded that a small percentage of the differences in SEMs
419 growth rate within individuals might be driven by underlying genetic background. These results
420 suggest other exposures may play a significant role, worthy of further investigation (39). Finally,
421 we showed using longitudinal data that once epimutations are established, most of them remain
422 stable in time. Previous findings suggested that methylation patterns are transmittable during cell
423 divisions(40). Given the above, we can speculate that SEMs could also be inherited through mitosis.
424 *SEMs association with cancer risk*
425 The main finding of the present study is the association of the number of SEMs with a
426 higher risk of breast and lung cancers and MBCN. The estimated ORs were not confounded by age
427 because we used age-matched case-control study design, and we further included age as adjustment
428 in logistic regression models. Further, the observed associations remained significant after
429 adjustment for smoking, BMI, physical activity, diet, alcohol consumption, and epigenetic clock
430 measures. Although in our study there is an association of the total number of SEMs with cancer

risk factors like smoking, obesity and epigenetic clocks, the results obtained in model 1 (minimally adjusted), model 2 (adjusted for various cancer risk factors), model 3 and model 4 (additionally adjusted for epigenetic clocks measures) did not differ significantly. The results above suggest that the increased number of SEMs consequence of unhealthy lifestyle explains a small part of the association of log(SEM) with cancer, meaning that other biological mechanisms are the main drivers of this associations. For example, endogenous exposures like inflammation or reduced DNA repair capacity (41) and other unmeasured environmental and lifestyle exposures (e.g. exposure to toxicants). In a manuscript currently under review from the MCCS group, they show that the DNAmGrimAge outperforms first-generation clocks in predicting different cancers, being the strongest association with lung cancer even after proper adjustment for smoking intensities and time. In this study the association of logSEM with breast and lung cancer and with MBCN remain significant after adjustment for DNAmGrimAge, suggesting SEMs and the new epigenetic clock as independent DNAm-based biomarkers, likely involving distinct biomolecular alterations. Further studies are needed to clarify better the underlying biological mechanisms linking SEMs and DNAmGrimAge to cancer.

Our results indicate that alterations of DNA methylation profiles could be detected in the blood years before cancer diagnosis, and together with previous studies, suggest that an increasing number of SEMs in blood could be predictive of risk of future cancers. The differences between cases and matched controls increased as the time from blood collection and cancer diagnosis decrease (**Figure S2**) in all but two types of cancer investigated, with a significant trend of increasing OR as the TTD decrease in breast, colorectal, prostate cancer and MBCN, further supporting the potential predictive utility of logSEM biomarker.

SEMs occur more likely in specific genomic regions

It is important to specify the meaning of the term ‘epimutation’: although some authors used this term in a broader sense (42), including epigenetic changes driven by DNA mutations, we are

referring to ‘epimutation’ as a switch of the ‘epigenetic state’ not due to underlying DNA sequence variations but to gain or loss of DNA methylation.

Our study suggests that regions and sites affected by epimutations are not entirely ‘stochastic’; instead, they are enriched in specific genomic regions, and randomly distributed inside them (34). This behaviour could be defined as ‘local, but non-global, stochasticity’. Our findings confirmed that epimutations preferentially occur in DNA sequences associated with open chromatin as previously observed by Ong et al.(43). Furthermore, SEMs were enriched in transcriptionally silenced genomic regions such as ‘inactive promoters’, ‘heterochromatin/low signal/CNV’, and ‘Polycomb-repressed’ regions. Additionally, epimutations more likely appear in TFBSs targeted by two members of PRC2: EZH2 and SUZ12, and the transcriptional corepressor ctBP2.

Consistently, smoking intensity was associated with enrichment of DNA methylation alterations in EZH2 and SUZ12 targets in buccal cells.(44). Similar patterns of DNAm alterations were described in normal breast tissue adjacent to cancerous breast tissue, compared with normal breast tissue in cancer-free women(45), and in our study comparing tumour with normal adjacent tissue using data from the TCGA project on breast, lung and colorectal cancer. Interestingly, EZH2-specific SEMs are significantly more stable in time (and conserved in tumour comparing with normal-adjacent tissue) compared with epimutations appearing in the rest of the genome.

SEMs in cancer tissue compared with adjacent normal tissue

To understand whether epimutation patterns in blood samples could be informative about epimutation patterns in the target tissue is crucial. Although DNAm from blood and tissue samples from the same individual are not available neither in our study nor in the databases available online, recent evidence suggests a strong correlation between DNAm profiles in blood and specific tissues (46,47). We analyzed epimutation profiles in DNAm data from tumours and normal adjacent tissue pairs from the TCGA project showing that the number of epimutations increased exponentially in tumour compared with normal adjacent tissue, as reported in previous studies using a slightly different analytical approach (45). In addition to previous studies, we showed that genomic regions

enriched of epimutations in both normal and tumour tissue are consistent with what observed in blood sample. Specifically, the enrichment of epimutations in TFBS of PRC2 complex is of particular interest, especially for its biological interpretation.

A *possible mechanism of carcinogenesis*

Being CpG sites targeted by EZH2 protein hypo-methylated in normal conditions (**Figure S5**), the vast majority of EZH2-specific SEMs appears as hypermethylation of a CpG site, suggesting crucial biomolecular mechanisms involved. The transcriptional regulation by DNA methylation and by PRC2 proteins are related: *in vitro* studies have demonstrated that they rarely act simultaneously on CpG islands(48), and removal of the epigenetic mark provokes a redistribution of the PRC2-distinctive H3K27me3 mark in mammalian cells. At a functional level, the link between aging, *PRC2* and global DNA methylation dysregulation involves the loss of self-renewal capacity of adult stem cells(49). Multipotent stem cell senescence *in vitro* is characterized by downregulation of *PRC2* genes, including *EZH2* and *SUZ12*.(49) Downregulation of *EZH2* and *SUZ12* may induce dysregulation of PRC2 targets, which include several tumour suppressor genes(50). For example, aberrant expression of *EZH2* was associated with alterations of *p53*, a known tumour suppressor gene(51).

The dynamics of the interaction between the Polycomb protein complex and DNA methylation are complex and not entirely understood. *In vitro* studies indicate that the two repressive systems are mutually exclusive and DNA methylation prevents Polycomb from accessing the promoter(52). The data reported in the present study suggests that aging may increase the enrichment of methylated sites in correspondence of TFBSs targeted by EZH2 and SUZ12, and consequently altering the efficacy of regulation of Polycomb. In line with these results, we could hypothesise that during aging, a more stable epigenetic silencing by DNA methylation could replace the plastic Polycomb repressive signal. Changes such as those described above might contribute to the early mechanisms involved in age-related diseases, specifically cancer. As proposed by other studies from Ohm et al.(53), Baylin et al.(54) and Widschwendter et al.(55) the tumour suppressive

genes regulated by Polycomb may switch from a dynamic to a fixed repressive state. In this context, tumour suppressor genes would not work properly, letting cells grow abnormally and become malignant. Vaz et al. suggested that these genes appear most vulnerable to aberrant promoter DNA methylation during cancer initiation and progression(56). More studies are needed to verify these data that raised new intriguing hypothesis connecting aging and cancer but the fact that SEMs data have been extracted from prospective study enforce previous studies done on cancer patients when the disease was already present (Tsai and Baylin, 2011 cell research).

Commented [GF6]: Ame, aggiungi citazione

Study limitations

Although most risk factors were measured carefully in the three cohort studies, the procedure to minimize variability due to the different sources of information possibly introduced bias in the regression models we used.

Besides, in the present study, we measured DNA methylation levels in blood and not in tissues. Tissue biopsy still represents the gold-standard approach for patients' diagnosis and prognostication. However, tissues do not represent tumour heterogeneity and, especially for early stages, residual disease and recurrence monitoring, a tissue biopsy sampling could be difficult or even dangerous (47). The evaluation of whole blood DNA methylation as a cancer risk marker is of particular interest because blood DNA constitutes a convenient 'tissue' to assay for constitutional methylation and its collection is non-invasive. Our results about SEMs using the TGCA data and recent literature suggest the methylation status of cancer tissues may reflects acquired or inherited somatic events that are detectable in non-targeted tissues (methylation memory of exposures/inheritance) and correlate with cancer susceptibility (46). Thus, epigenetic signatures in whole blood DNA could reflect the interaction of host genetic and environmental factors associated with cancer susceptibility as previously shown by others(57-59). Wong et al., for instance, showed that methylation of the BRCA1 promoter in blood DNA was more frequent in early-onset breast cancer patients and correlated with increased BRCA1 methylation levels in tumours(58). Finally, methylation in whole blood might reflect cancer predisposition as already demonstrated (60).

We found significant associations of SEMs with three out of eight cancers investigated and overall small magnitude in the effect sizes. This study results indicate accumulation of epimutations at a genome-wide level as a possible common biomarker in various cancers; however, each type of cancer is a well distinct disease, with its unique genetic landscape. The considerations above, indicate further research, possibly combining DNA methylation and gene expression data from both blood and tissue from the same individuals to understand better which specific genes or genomic regions influence cancer-risk when affected by SEMs, that is to investigate which epimutations are more deleterious than others. Future studies are also needed to identify cancer-specific epimutational signatures and to understand the biological mechanisms associated with accumulation of epimutations during the lifespan, possibly involving genetic background and DNA-repair capacity.

Conclusions

To our knowledge, this is the most extensive study on the association of SEMs with cancer risk using a prospective study design. A higher number of SEMs was significantly associated with an increased risk of breast and lung cancer and with MBCN. Also, we confirmed previous observation about the exponential increase of SEMs during aging using longitudinal data, showing that most of SEMs are stable in time and conserved in tumour compared with normal-adjacent tissue. Finally, we showed that SEMs more likely occur in specific genomic regions, suggesting a biomolecular mechanism involving PRC2 proteins, which may deserve further investigation. If confirmed with additional studies *in vitro*, these observations might open new avenues for the understanding of carcinogenesis biomolecular mechanisms.

Acknowledgments

The Authors are very thankful to Dr Akram Ghantous (IARC, Lyon, France) for the methylation analyses of PEM-Turin study, produced within the Exposomics EC FP7 grant (Grant agreement no: 308610 to PV). The results here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

560 **Figure legends**

561 **Figure 1.** Exponential increase of the total number of SEMs with age: mean and 95% confidence
562 interval of the total number of SEMs (on a logarithmic scale) by age group in the three study
563 cohorts, in cases and controls combined (top) and in controls only (bottom). R and p-values refer to
564 Pearson Correlation test.

565 **Figure 2.** Total number of SEMs and risk of breast and lung cancer. Forest plots representing the
566 three-studies random effect (RE) maximum likelihood meta-analysis for breast (A) and lung cancer
567 (B), and the meta-analysis of EPIC and MCCS for colorectal cancer (C).

568 **Supplementary figure legends**

569 **Figure S1.** Lack of correlation between log(SEM) and white blood cells (WBC) proportions:
570 heatmap of Pearson correlation coefficients including log(SEM) and WBC proportions estimated
571 using Houseman algorithm.

572 **Figure S2** Odds ratio (ORs) significantly increase as TTD decrease in breast, colorectal, prostate
573 cancer and MBCN: Forest plots indicating ORs stratified by the time-to-disease and type of cancer.
574 P-values refer to the Cochran Armitage test for trend.

575 **Figure S3** Total number of SEMs is associated with Horvath DNAmAge epigenetic clock:
576 Scatterplots of log(SEM) on the x-axis and DNAmAge on the y-axis, in EPIC (A), MCCS (B) and
577 NOWAC (C) (cases and controls combined on the top, controls only on the bottom). P-values refer
578 to the Pearson correlation test.

579 **Figure S4.** Total number of SEMs is associated with DNAmGrimAge epigenetic clock: Scatterplots
580 of log(SEM) on the x-axis and DNAmGrimAge on the y-axis, in EPIC (A), MCCS (B) and
581 NOWAC (C) (cases and controls combined on the top, controls only on the bottom). P-values refer
582 to the Pearson correlation test.

583 **Figure S5.** The number of EZH2-specific SEMs correlates with the total number of SEMs genome-
584 wide: Scatterplots of log(SEM) genome-wide on the x-axis and EZH2-specific logSEM on the y-

axis in EPIC (A), MCCS (B) and NOWAC (C) (cases and controls combined on the top, controls only on the bottom). P-values refer to Spearman correlation tests.

Figure S6. The majority of CpG sites targeted by EZH2 are on average hypomethylated: Histogram of average DNAm values for the CpGs targeted by EZH2 protein.

Figure S7. The total number of SEMs in the PEM-Turin dataset significantly increase over time: Spaghetti plot showing the increasing trend of log(SEM) over time. Each line indicates a single individual in the PEM-Turin dataset.

Figure S8. SEMs exponentially increase in tumour compared with normal-adjacent tissue: boxplot of log(SEM) in normal and tumor tissue of lung (A), breast (B) and colorectal cancer (C) (genome-wide logSEM on the top, EZH2-specific logSEM on the bottom). These data come from the TCGA project.

Figure S9. Batch effect does not influence logSEM computation: Scatterplots for the association of logSEM with batch adjusted logSEM in EPIC (A), MCCS (B) and NOWAC (C). P-values refer to Pearson correlation tests.

References

1. Ben-Shlomo Y, Kuh D. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. Int J Epidemiol **2002**;31(2):285-93.
2. Berger NA, Savvides P, Koroukian SM, Kahana EF, Deimling GT, Rose JH, *et al.* Cancer in the elderly. Transactions of the American Clinical and Climatological Association **2006**;117:147-55; discussion 55-6.
3. Lund E. An exposure driven functional model of carcinogenesis. Med Hypotheses **2011**;77(2):195-8 doi 10.1016/j.mehy.2011.04.009.
4. Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. The hallmarks of aging. Cell **2013**;153(6):1194-217 doi 10.1016/j.cell.2013.05.039.
5. Rozhok AI, DeGregori J. The evolution of lifespan and age-dependent cancer risk. Trends Cancer **2016**;2(10):552-60 doi 10.1016/j.trecan.2016.09.004.
6. Jones MJ, Goodman SJ, Kobor MS. DNA methylation and healthy human aging. Aging Cell **2015**;14(6):924-32 doi 10.1111/acel.12349.
7. Zheng SC, Widschwendter M, Teschendorff AE. Epigenetic drift, epigenetic clocks and cancer risk. Epigenomics **2016**;8(5):705-19 doi 10.2217/epi-2015-0017.
8. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol **2013**;14(10):R115 doi 10.1186/gb-2013-14-10-r115.
9. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell **2013**;49(2):359-67 doi 10.1016/j.molcel.2012.10.016.
10. Dugue PA, Bassett JK, Joo JE, Baglietto L, Jung CH, Wong EM, *et al.* Association of DNA Methylation-Based Biological Age With Health Risk Factors and Overall and Cause-Specific Mortality. Am J Epidemiol **2018**;187(3):529-38 doi 10.1093/aje/kwx291.
11. Dugue PA, Bassett JK, Joo JE, Jung CH, Ming Wong E, Moreno-Betancur M, *et al.* DNA methylation-based biological aging and cancer risk and survival: Pooled analysis of seven prospective studies. Int J Cancer **2018**;142(8):1611-9 doi 10.1002/ijc.31189.
12. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, *et al.* DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol **2015**;16:25 doi 10.1186/s13059-015-0584-6.
13. Chen BH, Marioni RE, Colicino E, Peters MJ, Ward-Caviness CK, Tsai PC, *et al.* DNA methylation-based measures of biological age: meta-analysis predicting time to death. Aging (Albany NY) **2016**;8(9):1844-65 doi 10.18632/aging.101020.
14. Horvath S, Gurven M, Levine ME, Trumble BC, Kaplan H, Allayee H, *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol **2016**;17(1):171 doi 10.1186/s13059-016-1030-0.
15. Marioni RE, Shah S, McRae AF, Ritchie SJ, Muniz-Terrera G, Harris SE, *et al.* The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. Int J Epidemiol **2015**;44(4):1388-96 doi 10.1093/ije/dyu277.
16. Horvath S, Erhart W, Brosch M, Ammerpohl O, von Schonfels W, Ahrens M, *et al.* Obesity accelerates epigenetic aging of human liver. Proc Natl Acad Sci U S A **2014**;111(43):15538-43 doi 10.1073/pnas.1412759111.
17. Quach A, Levine ME, Tanaka T, Lu AT, Chen BH, Ferrucci L, *et al.* Epigenetic clock analysis of diet, exercise, education, and lifestyle factors. Aging (Albany NY) **2017**;9(2):419-46 doi 10.18632/aging.101168.
18. Fiorito G, Polidoro S, Dugue PA, Kivimaki M, Ponzi E, Matullo G, *et al.* Social adversity and epigenetic aging: a multi-cohort study on socioeconomic differences in peripheral blood DNA methylation. Sci Rep **2017**;7(1):16266 doi 10.1038/s41598-017-16391-5.

19. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, *et al.* An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) **2018**;10(4):573-91 doi 10.18632/aging.101414.
20. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging (Albany NY) **2019**;11(2):303-27 doi 10.18632/aging.101684.
21. Dugué P, Li S, Hopper JL, Milne RL. Chapter 3 - DNA Methylation-Based Measures of Biological Aging. Academic Press; 2018.
22. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet **2013**;22(R1):R7-R15 doi 10.1093/hmg/ddt375.
23. Yamashita S, Kishino T, Takahashi T, Shimazu T, Charvat H, Kakugawa Y, *et al.* Genetic and epigenetic alterations in normal tissues have differential impacts on cancer risk among tissues. Proc Natl Acad Sci U S A **2018**;115(6):1328-33 doi 10.1073/pnas.1717340115.
24. Haque MM, Nilsson EE, Holder LB, Skinner MK. Genomic Clustering of differential DNA methylated regions (epimutations) associated with the epigenetic transgenerational inheritance of disease and phenotypic variation. BMC Genomics **2016**;17:418 doi 10.1186/s12864-016-2748-5.
25. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, *et al.* Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging (Albany NY) **2019**;11(7):2045-70 doi 10.18632/aging.101900.
26. Gentilini D, Scala S, Gaudenzi G, Garagnani P, Capri M, Cescon M, *et al.* Epigenome-wide association study in hepatocellular carcinoma: Identification of stochastic epigenetic mutations through an innovative statistical approach. Oncotarget **2017**;8(26):41890-902 doi 10.18632/oncotarget.17462.
27. Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, Kitchener HC, *et al.* Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. Genome Med **2012**;4(3):24 doi 10.1186/gm323.
28. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun **2015**;6:10192 doi 10.1038/ncomms10192.
29. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung CH, *et al.* DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer **2017**;140(1):50-61 doi 10.1002/ijc.30431.
30. van Veldhoven K, Polidoro S, Baglietto L, Severi G, Sacerdote C, Panico S, *et al.* Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clin Epigenetics **2015**;7:67 doi 10.1186/s13148-015-0104-2.
31. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics **2012**;13:86 doi 10.1186/1471-2105-13-86.
32. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics **2017**;18(1):105 doi 10.1186/s12859-017-1511-5.
33. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, *et al.* Missing value estimation methods for DNA microarrays. Bioinformatics **2001**;17(6):520-5 doi 10.1093/bioinformatics/17.6.520.
34. Gentilini D, Garagnani P, Pisoni S, Bacalini MG, Calzari L, Mari D, *et al.* Stochastic epigenetic mutations (DNA methylation) increase exponentially in human aging and correlate with X chromosome inactivation skewing in females. Aging (Albany NY) **2015**;7(8):568-78 doi 10.18632/aging.100792.
35. Breusch TS. Maximum likelihood estimation of random effects model. Journal of econometrics **1987**;36(3):383-9 doi 10.1016/0304-4076(87)90010-8.

36. Viechtbauer W. Conducting Meta-Analyses in R with the metafor Package. Journal of Statistical Software **2010**;36(3).
37. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, *et al.* The exposome in practice: Design of the EXPOsOMICS project. Int J Hyg Environ Health **2017**;220(2 Pt A):142-51 doi 10.1016/j.ijheh.2016.08.001.
38. Mancini FR, Laine JE, Tarallo S, Vlaanderen J, Vermeulen R, van Nunen E, *et al.* microRNA expression profiles and personal monitoring of exposure to particulate matter. Environ Pollut **2020**;263(Pt B):114392 doi 10.1016/j.envpol.2020.114392.
39. Wang Y, Karlsson R, Jylhava J, Hedman AK, Almqvist C, Karlsson IK, *et al.* Comprehensive longitudinal study of epigenetic mutations in aging. Clin Epigenetics **2019**;11(1):187 doi 10.1186/s13148-019-0788-9.
40. Robertson KD. DNA methylation, methyltransferases, and cancer. Oncogene **2001**;20(24):3139-55 doi 10.1038/sj.onc.1204341.
41. Slyskova J, Korenkova V, Collins AR, Prochazka P, Vodickova L, Svec J, *et al.* Functional, genetic, and epigenetic aspects of base and nucleotide excision repair in colorectal carcinomas. Clin Cancer Res **2012**;18(21):5878-87 doi 10.1158/1078-0432.CCR-12-1380.
42. Oey H, Whitelaw E. On the meaning of the word 'epimutation'. Trends Genet **2014**;30(12):519-20 doi 10.1016/j.tig.2014.08.005.
43. Ong ML, Holbrook JD. Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. Aging Cell **2014**;13(1):142-55 doi 10.1111/acel.12159.
44. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao Y, Jones A, *et al.* Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. JAMA Oncol **2015**;1(4):476-85 doi 10.1001/jamaoncol.2015.1053.
45. Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nat Commun **2016**;7:10478 doi 10.1038/ncomms10478.
46. Tahara T, Maegawa S, Chung W, Garriga J, Jelinek J, Estecio MR, *et al.* Examination of whole blood DNA methylation as a potential risk marker for gastric cancer. Cancer Prev Res (Phila) **2013**;6(10):1093-100 doi 10.1158/1940-6207.CAPR-13-0034.
47. Constancio V, Nunes SP, Henrique R, Jeronimo C. DNA Methylation-Based Testing in Liquid Biopsies as Detection and Prognostic Biomarkers for the Four Major Cancer Types. Cells **2020**;9(3) doi 10.3390/cells9030624.
48. Brinkman AB, Gu H, Bartels SJ, Zhang Y, Matarese F, Simmer F, *et al.* Sequential ChIP-bisulfite sequencing enables direct genome-scale investigation of chromatin and DNA methylation cross-talk. Genome Res **2012**;22(6):1128-38 doi 10.1101/gr.133728.111.
49. Jung JW, Lee S, Seo MS, Park SB, Kurtz A, Kang SK, *et al.* Histone deacetylase controls adult stem cell aging by balancing the expression of polycomb genes and jumonji domain containing 3. Cell Mol Life Sci **2010**;67(7):1165-76 doi 10.1007/s00018-009-0242-9.
50. Zingg D, Debbache J, Schaefer SM, Tuncer E, Frommel SC, Cheng P, *et al.* The epigenetic modifier EZH2 controls melanoma growth and metastasis through silencing of distinct tumour suppressors. Nat Commun **2015**;6:6051 doi 10.1038/ncomms7051.
51. Shiogama S, Yoshiba S, Soga D, Motohashi H, Shintani S. Aberrant expression of EZH2 is associated with pathological findings and P53 alteration. Anticancer Res **2013**;33(10):4309-17.
52. Sproul D, Meehan RR. Genomic insights into cancer-associated aberrant CpG island hypermethylation. Brief Funct Genomics **2013**;12(3):174-90 doi 10.1093/bfgp/els063.
53. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, *et al.* A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet **2007**;39(2):237-42 doi 10.1038/ng1972.

54. Baylin SB, Ohm JE. Epigenetic gene silencing in cancer - a mechanism for early oncogenic pathway addiction? Nat Rev Cancer **2006**;6(2):107-16 doi 10.1038/nrc1799.
55. Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, *et al.* Epigenetic stem cell signature in cancer. Nat Genet **2007**;39(2):157-8 doi 10.1038/ng1941.
56. Vaz M, Hwang SY, Kagiampakis I, Phallen J, Patil A, O'Hagan HM, *et al.* Chronic Cigarette Smoke-Induced Epigenomic Changes Precede Sensitization of Bronchial Epithelial Cells to Single-Step Transformation by KRAS Mutations. Cancer Cell **2017**;32(3):360-76 e6 doi 10.1016/j.ccell.2017.08.006.
57. Marsit CJ, Koestler DC, Christensen BC, Karagas MR, Houseman EA, Kelsey KT. DNA methylation array analysis identifies profiles of blood-derived DNA methylation associated with bladder cancer. J Clin Oncol **2011**;29(9):1133-9 doi 10.1200/JCO.2010.31.3577.
58. Wong EM, Southey MC, Fox SB, Brown MA, Dowty JG, Jenkins MA, *et al.* Constitutional methylation of the BRCA1 promoter is specifically associated with BRCA1 mutation-associated pathology in early-onset breast cancer. Cancer Prev Res (Phila) **2011**;4(1):23-33 doi 10.1158/1940-6207.CAPR-10-0212.
59. Brennan K, Garcia-Closas M, Orr N, Fletcher O, Jones M, Ashworth A, *et al.* Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. Cancer Res **2012**;72(9):2304-13 doi 10.1158/0008-5472.CAN-11-3157.
60. Hao X, Luo H, Krawczyk M, Wei W, Wang W, Wang J, *et al.* DNA methylation markers for diagnosis and prognosis of common cancers. Proc Natl Acad Sci U S A **2017**;114(28):7414-9 doi 10.1073/pnas.1703577114.