Bio-statistical approaches to evaluate the link between specific nutrients and methylation patterns in a breast cancer case-control study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) study

Flavie Perrier

N°d'ordre NNT : 2018LYSE1146

THESE de DOCTORAT DE L'UNIVERSITE DE LYON

opérée au sein de

l'Université Claude Bernard Lyon 1

Ecole Doctorale N°205

Interdisciplinaire Sciences Santé (EDISS)

Spécialité de doctorat

Epidémiologie, santé publique, recherche sur les services de santé

Discipline

Biostatistiques

Soutenue publiquement le 13/09/2018, par :

Flavie PERRIER

Bio-statistical approaches to evaluate the link between specific nutrients and methylation patterns in a breast cancer case-control study nested within the European Prospective Investigation into Cancer and Nutrition (EPIC) study

---

Devant le jury composé de :

RONDEAU, Virginie, PhD ***Présidente, Rapporteur***
Directrice de Recherche, INSERM CR1219 – ISPED, Bordeaux (France).

VERMEULEN, Roel, PhD ***Rapporteur***
Visting professor, Department of Epidemiology and Biostatistics, Imperial College London, London (UK).
Associate Professor, Environmental Epidemiology, Institute for Risk Assessment Sciences (IRAS), Utrecht University, Utrecht (The Netherlands).
Adjunct Professor, Molecular Epidemiology, Julius Center, University Medical Center Utrecht, Utrech (The Netherlands).

POLIDORO, Silvia, PhD ***Examinatrice***
Senior researcher, Italian Institute for Genomic Medicine (IIGM), Torino, (Italy).

VIALLON, Vivian, PhD ***Examinateur***
Maitre de conférences, Centre International de Recherche sur le Cancer, Lyon (France).

FERRARI, Pietro, PhD ***Directeur de thèse***
Directeur d'équipe, Centre International de Recherche sur le Cancer, Lyon (France).

ROMIEU, Isabelle, PhD ***Invitée***
Centre International de Recherche sur le Cancer, Lyon (France).

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. le Professeur Frédéric FLEURY**
Président du Conseil Académique	M. le Professeur Hamda BEN HADID
Vice-président du Conseil d'Administration	M. le Professeur Didier REVEL
Vice-président du Conseil Formation et Vie Universitaire	M. le Professeur Philippe CHEVALIER
Vice-président de la Commission Recherche	M. Fabrice VALLÉE
Directrice Générale des Services	Mme Dominique MARCHAND

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur G.RODE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. X. PERROT
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : Mme la Professeure A-M. SCHOTT

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. THEVENARD
Département Chimie Biochimie	Directeur : Mme C. FELIX
Département GEP	Directeur : M. Hassan HAMMOURI
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur G. TOMANOV
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : M. le Professeur J-C PLENET
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. Y.VANPOULLE
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. le Professeur E.PERRIN
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. VITON
Ecole Supérieure du Professorat et de l'Education	Directeur : M. le Professeur A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Directeur : M. N. LEBOISNE

RESUME

De par les centaines de milliers de données qui les caractérisent, les bases de données épigénétiques représentent actuellement un défi majeur. L'objectif principal de cette thèse est d'évaluer la performance d'outils statistiques développés pour les données de grande dimension, en explorant l'association entre facteurs alimentaires reliés au cancer du sein (CS) et méthylation de l'ADN dans la cohorte EPIC.

Afin d'étudier les caractéristiques des données de méthylation, l'identification des sources systématiques de variabilité des mesures de méthylation a été effectuée par la méthode de la PC-PR2. Ainsi la performance de trois techniques de normalisation, très répandues pour corriger la part de variabilité non désirée, a été évaluée en quantifiant l'entendu de variabilité attribuée aux facteurs de laboratoire avant et après chaque méthode de correction.

Une fois la méthode de normalisation la plus appropriée identifiée, la relation entre le folate, l'alcool et la méthylation de l'ADN a été analysée par le biais de trois approches : une analyse individuelle des sites CpG, une analyse de DMR et la régression fused lasso. Les deux dernières méthodes visent à identifier des régions spécifiques de l'épigénome grâce aux corrélations possibles entre les sites proches. La méthylation globale a aussi été utilisée pour étudier la relation entre méthylation et risque de CS.

Grâce à une évaluation exhaustive d'outils statistiques révélant la complexité des données de méthylation de l'ADN, cette thèse offre un aperçu instructif de connaissances pour les études épigénétiques, avec une possibilité d'application de méthodologie similaire aux analyses d'autres types de données -*omiques*.

Mots-clés : Epigénétique, PC-PR2, méthylation, DMR, fused lasso, cancer du sein, EPIC.

Institut de préparation de la thèse :
Centre International de Recherche sur le Cancer (CIRC),
Groupe de Méthodologie Nutritionnelle et Biostatistique, Section Nutrition et Métabolisme,
150 cours Albert Thomas 69372 Lyon Cedex 08, France.

ABSTRACT

Epigenetics data are challenging sets characterized by hundreds of thousands of features. The main objective of this thesis was to evaluate the performance of some of the existing statistical methods to handle sets of large dimension data, exploring the association between dietary factors related to breast cancer (BC) and DNA methylation within the EPIC study.

In order to investigate the characteristics of epigenetics data, the identification of random and systematic sources of variability of methylation measurements was attempted, via the principal component partial R-square (PC-PR2) method. Using this technique, the performance of three popular normalization techniques to correct for unwanted sources of variability was evaluated by quantifying epigenetics variability attributed to laboratory factors before and after the application of each correction method.

Once a suitable normalization procedure was identified, the association between alcohol intake, dietary folate and methylation levels was examined by means of three approaches: an analysis of individual CpG sites, of differentially methylated regions (DMRs) and using fused lasso regression. The last two methods aim at the identification of specific regions of the epigenome using the potential correlation between neighboring CpG sites. Global methylation levels were used to investigate the relationship between methylation and BC risk.

By performing an exhaustive evaluation of the statistical tools used to disclose complexity of DNA methylation data, this thesis provides informative insights for studies focusing on epigenetics, with promising potentials to apply similar methodology to the analysis of other *-omics* data.

Keywords: Epigenetics, PC-PR2, methylation, DMR, fused lasso, breast cancer, EPIC.

Institute hosting the thesis candidate:
International Agency of Research on Cancer (IARC),
Nutritional Methodology and Biostatistics group, Nutrition and Metabolism section,
150 cours Albert Thomas 69372 Lyon Cedex 08, France.

RESUME SUBSTANTIEL

Les nouvelles avancées technologiques dans le domaine *-omiques* rendent possible l'acquisition de plus en plus de données par individu, allant de quelques centaines jusqu'à plusieurs milliers. Un des défis les plus importants actuellement engendré par ces données est de surmonter les contraintes liées à leur très grande dimension. Habituellement, en épidémiologie, le nombre de facteurs étudiés est inférieur au nombre de participants de la population d'étude. Cependant, en présence de données *-omiques*, telles que les données épigénétiques caractérisées par des centaines de milliers de mesures par individu, le nombre de facteurs étudiés est nettement supérieur au nombre de participants dans la population. Les méthodes statistiques usuellement utilisées en épidémiologie ne sont alors plus nécessairement adaptées à ces données. L'objectif principal de cette thèse est d'évaluer la performance d'outils statistiques développés pour les données de grande dimension, en utilisant comme exemple, l'association entre certains facteurs alimentaires reliés au cancer du sein (CS) et la méthylation de l'ADN dans la cohorte européenne prospective EPIC (European Prospective Investigation into Cancer and nutrition).

La méthylation de l'ADN est altérée par de nombreux facteurs incluant l'âge et des facteurs environnementaux tels que la consommation d'alcool et de tabac. En plus de ces facteurs, des variabilités systématiques et aléatoires peuvent aussi être introduites lors du traitement technique des échantillons biologiques tels que le "batch" (i.e. groupe d'échantillons traités en même temps) ou la position des échantillons à l'intérieur du "batch". De plus dans le cas d'une cohorte multicentrique comme EPIC, le centre dont sont issues les données peut également engendrer de la variabilité due à une collecte et un traitement des échantillons pouvant varier entre les centres. Toutes ces variabilités peuvent compromettre la justesse du procédé de mesure de la méthylation et biaiser l'estimation des associations investiguées. Afin de mieux appréhender la complexité des données de méthylation de l'ADN, l'identification des sources systématiques et aléatoires de variabilité introduites pendant l'acquisition des mesures de méthylation est nécessaire. En se servant de la méthode "principal component partial R-square" (PC-PR2) qui combine une technique de réduction de dimensions (analyse en composantes principales) avec une modélisation de régression linéaire, trois techniques déjà existantes, développées pour corriger les données de méthylation pour des facteurs de variabilité, ont été comparées : ComBat, SVA et une méthode de régression pour le calcul de résidus. Avant et après application de chacune des trois techniques de normalisation, la méthode de la PC-PR2 a été utilisée afin de quantifier la part de variabilité de chaque facteur lié au traitement des échantillons. Les trois méthodes ont réussi à enlever la part de variabilité attribuée au traitement des échantillons. Parmi les

trois méthodes testées, SVA s'est avérée être la méthode produisant les résultats les plus conservatifs dans une application visant à comparer l'association entre le statut tabagique et la méthylation de l'ADN.

L'alcool et le folate sont connus pour être, respectivement, positivement et inversement associés au risque de CS. Leurs effets antagonistes sont également reconnus dans le métabolisme monocarboné (OCM) qui est essentiel pour la réplication et la réparation de l'ADN. En diminuant l'absorption de folate, en augmentant son excrétion par les reins et en inhibant la synthase de la méthionine, l'alcool peut entrainer un dysfonctionnement de l'OCM, ce qui pourrait amener à une synthèse anormale de l'ADN et donc impacter sur le risque de CS. Afin d'étudier l'association entre l'apport alimentaire en folate et la consommation d'alcool avec la méthylation de l'ADN, trois méthodes statistiques ont été utilisées. La première méthode a analysé l'association du folate et de l'alcool sur la méthylation de l'ADN séparément pour chaque site CpG, alors que les analyses de DMR (differentially methylated region) et de fused lasso (FL) avaient pour but d'identifier des régions spécifiques de l'épigénome. Une faible association entre la consommation d'alcool et le niveau de méthylation de deux sites CpG a été observée. Les résultats des analyses de DMRs et de FL ont montré que le folate et l'alcool étaient associés avec des altérations du niveau de méthylation dans certaines régions de l'épigénome, dont certaines sont associées avec des gènes connus pour leur rôle de suppresseurs de tumeurs tels que les gènes *GSDMD* et *HOXA5*. Ces résultats sont en accord avec l'hypothèse supportant l'idée que des mécanismes épigénétiques pourraient avoir un rôle dans l'association entre folate, alcool et le risque de CS.

La méthylation de l'ADN est suspectée d'être impliquée dans le développement du CS par le biais de dysfonctionnements de mécanismes cellulaires. Cependant, pour le moment aucune association entre méthylation individuelle de site CpG et risque de CS n'a été validée. Seule une association positive entre hypo-méthylation globale et CS a été observée de façon récurrente au sein des études prospectives d'association à l'échelle de l'épigénome (EWAS). La méthylation globale de l'ADN, définie comme la moyenne des niveaux de méthylation de l'ensemble des sites CpG, a été évaluée au sein d'une étude coordonnée par le groupe d'Epigénétique du CIRC par rapport au risque de CS. Les résultats des analyses statistiques ont révélé une faible association positive entre la méthylation moyenne des sites appartement à un îlot de CpG sites et le risque de CS.

Grâce à une évaluation exhaustive d'outils statistiques révélant la complexité des données de méthylation de l'ADN, cette thèse offre un aperçu instructif de connaissances pour les études des données épigénétiques. La méthodologie présentée dans cette thèse ouvre aussi

la possibilité à des applications similaires adaptées aux analyses statistiques d'autres types de données -*omiques*.

ACKNOWLEDGEMENTS

I would like to gratefully acknowledge Isabelle Romieu who has initiated this project and co-supervised me during these last four years of PhD. It started for me four years ago when Isabelle presented her work on breast cancer risk in the institute where I was doing my master internship. I am really thankful she accepted to have a quick talk with me that day. This 5 minutes unexpected meeting initiated a very instructive and enriching collaboration.

I would like to express my sincere gratitude to Pietro Ferrari who co-supervised me during this thesis. I am truly thankful he always encouraged me and helped me to bring out the best in me.

I would like to thank Vivian Viallon for the statistical and mathematical expertise he kindly shared with me.

I would like to thank Véronique Chajès for her very useful advices and support.

I am warmly grateful for Béatrice Fervers, David Cox and Marc Chadeau-Hyam for the precious advices they provided in their different areas during the annual thesis follow-up committees.

I would like to acknowledge my colleagues from the Epigenetics group, and especially Zdenko Herceg, Srikant Ambatipudi, Akram Ghantous and Hector Hernandez-Vargas who were of precious help, particularly to better understand the specificity of the epigenetics area. At the beginning the communication was not easy between statisticians and biologists. We rapidly noticed that the technical languages used were sometimes different to refer to same terms. Thanks to them, we finally agreed on a very useful common language.

I would like to thank la Fondation de France and the French National Cancer Institute (INCa) who have financed this work.

I would like to express my sincere gratitude to all my colleagues from the Nutrition and Metabolism Section, and especially from the NMB group, for their support and help during the last four years. I am warmly thankful to have the occasion to work and share daily great moment with them. I had the chance to meet so many great persons with so many different scientific and cultural backgrounds. I am especially grateful to: Amina, Anne-Sophie, Benhaz, Benjamin, Claudia, Elom, Emilie, Hwayoung, Karina, Kayo, Kuanrong, Laura, Lola, Magda, Manon, Marco, Marta, Michèle, Minkyung, Nada, Pauline, Rachel, Sémi, Sabine, Talita, Tristan. I have learned something from each of them. They all participated in a way to this work, even though a short chat around a cup of coffee or tea. This international networking is one of the important strength of IARC.

ABBREVIATIONS

BC: breast cancer;

BMI: body mass index;

CI: confidence interval;

CpG: cytosine-phosphate-guanine;

DMR: differentially methylated region;

EPIC: European Prospective Investigation into Cancer and nutrition;

ER: estrogen receptor;

FDR: false discovery rate;

FL: fused lasso;

HER2: human epidermal growth factor 2;

HM450K: Illumina Infinium HumanMethylation450K BeadChip;

IEAA: intrinsic epigenetic age accelerating;

LASSO: least absolute shrinkage and selection operator;

OCM: one-carbon metabolism;

OLS: ordinary least squares;

OR: odd ratio;

PC: principal component;

PC-PR2: principal component partial R-square;

PCA: principal component analysis;

PCR: principal component regression;

PLS: partials least squares;

PR: progesterone receptor;

SAM: S-adenosyl methionine;

SD: standard deviation;

SVA: surrogate variable analysis;

WCRF: World Cancer Research Fund.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

INTRODUCTION

In the era of *-omics*, large amount of data are generated in epidemiological investigations by new generation of high-throughput acquisition platforms on large number of biological features, such as epigenetics, metabolomics, transcriptomics, proteomics, etc. This novel context generates a number of new issues related to the management, the characterization and the analysis of very complex sets of data. Traditionally in epidemiological studies, the numbers of exposure variables (p) is lower than the sample size (n). Classical statistical methods typically require the size of the study population to be large and a multiple of the number of variables. In *-omics,* high-throughput datasets characterized by large number of variables, the number of subjects might be limited due to technical and economical limitations of the experiment. In this case, the number of features can be way larger than the sample size, a situation classically known as $p \gg n$.

Standard methods to analyze *-omics* data generally involve the use of specific statistical techniques, such as regression modeling, complemented by methods to account for multiple testing, such as the false discovery rate (FDR) or Bonferroni corrections. The method is relatively straightforward to implement but, as for *-omics* data, p may reach the range of hundreds of thousands variables, and FDR and Bonferroni solutions are very demanding in terms of statistical power to preserve a nominal level of statistical significance. This increases the likelihood of capturing medium-to-large associations, but leaves little margin to focus the investigation effort on numerous, potentially relevant, weak associations. In addition, *-omics* data reflects the complexity of biological systems expressing a multitude of features related to metabolism, genetic and protein profiles, changes in gene expression, acquired from biological samples (urine, blood, saliva, tissues). As a results, datasets often have unknown structures, and the little is known on the way these features interact in response to environmental exposures. The high dimensionality of this data, coupled with their biological complexity make the application of classical statistical tools for research purposes not straightforward. Novel statistical tools have been recently proposed in the scientific literature to fully exploit the potential of a wealth of new data, either by conceiving new statistical techniques or by re-adapting existing tools to *-omics* analyses (1-3).

Statistical methods for large dimension data

Recent progress in technology made it possible the acquisition of thousands of features for relatively sizeable amount of study participants' samples, typically from few tens to several hundreds. This situation generated the need of conceiving solutions for the process of numerous samples in sequence. A standard way to handle large volumes of samples with a

limited number of machines (very often just one) is to allocate samples in laboratory batches, which allows the process of a group of samples at same time (4). Within a batch, samples might be separated into several chips, also referred as arrays. Due to technical limitations of the machines, only a limited number of samples can be handled in a same batch; thus several batches are usually needed to process all the samples. It is unlikely that all the batches were processed with the exact same experimental conditions: several technicians may handle or prepare the samples; the room temperature during samples processing may also change, etc. These differences may introduce variability in the features measurement. The 'batch effect', has been documented in the scientific literature (4, 5), and it is not the only source of systematic variability introduced by technical processing of samples. A 'positional effect', i.e. the physical position of samples on the chip, has also been observed (6). Unwanted biological variation can be a problem as well. Some factors may introduce systematic variation i.e. that affects all samples from a group in a similar manner whereas others introduce variation, which can be assumed to be random, namely caused by unpredictable or uncertain factors. As a result, technical management of samples likely introduces unwanted technical variability in *-omics* measurements that might compromise the accuracy of the measurement process and introduce bias in the estimation process of the association of interest. Careful random allocation of samples over chips (7, 8) is essential to make it independent from specific characteristics of the samples, i.e. country of origin, BMI, age. As a result, random and systematic technical variability need to be addressed. Some correction methods suggested in the literature require an *a priori* identification of factors potentially influencing variation (9-11). The large dimension of *-omics* data makes it difficult to quantify the amount of variability attributable to sources of systematic and random variation. The principal component partial R-square (PC-PR2) method was developed to quantify systematic and random variation in metabolomics data (3). The method is based on the combination of principal component analysis (PCA), which summarized the information given by a set of features in a reduced number of components that maximize the variance in the feature matrix, with the concept of the partial R^2 statistics in multivariable linear regression. A particularly appealing feature of the method is the capacity of successfully performing PCA in presence of hundreds to thousands of features. The technique could be extended to other *-omics* data.

Once major sources of systematic and random variability have been identified, another challenge is the treatment of unwanted variability data among a wealth of normalization techniques proposed in the literature. A popular way to tackle this would involve the computation of residuals from regression model where the outcome is, in turn, each feature from the dataset and the predictors covariates are the factors identified as expressing the

major sources of variability. Normalization techniques are usually specific to the *-omics* set under investigation. For example, the most popular techniques for DNA methylation data are the Surrogate Variables Analysis (SVA) (12, 13) and the ComBat technique (9). SVA is a method developed to remove variability originating from pre-identified factors but also unknown sources, through the estimation of surrogate variables potentially influencing overall variability. The ComBat method is a procedure based on an empirical Bayes approach with an additive and a multiplicative component, the latter contributing to shrink the feature-specific variability, thus also handling outlier values.

After accounting for unwanted variability, the data can be analyzed following diverse approaches. Standard analyses, involving the evaluation of each feature separately, can be complemented by techniques to handle several independent covariates in the linear predictor. The ordinary least square (OLS) method estimates coefficients from a linear multivariable regression model by minimizing the sum of the squares of the error terms. In regression analyses, over-fitting may occur when the number of predictors exceeds 10% of the number of observations. In addition, collinearity that inflates parameter estimates' variability may occur when there are many predictors and the model may also be difficult to interpret. Moreover the model is no longer identifiable when the number of predictors exceeds the number of observations, as for *-omics* analyses. A first solution would be to use a penalized approach such as Ridge regression (14), the Lasso (15) or elastic net (16), which introduce penalties in the OLS fit function to control the trade-off between goodness of fit and the number of predictors, an element referred to as model complexity. The penalty introduced in Ridge regression improves prediction error by shrinking large regression coefficients, but it does not reduce the model complexity. The Lasso imposes a penalty to encourage sparsity of coefficients, i.e. by setting to null coefficients, thus achieving shrinkage of parameter estimates and variable selection simultaneously. However, it can only select at most n variables out of p candidates. Elastic net combines Ridge and Lasso penalties, and it can be viewed as a compromise between the two approaches. Elastic net is particularly useful when the number of predictors (p) is much larger than the number of observations (n).

Instead of performing features selection, other statistical methods aim at reducing the dimension of the features set while keeping most of its variability. Principal component analysis (PCA) is a dimension reduction technique that constructs orthogonal principal components (PC) defined as linear combinations of the original features with maximal variance. The original set of correlated features is converted into a set of linearly independent variables. The PCs can be used as predictors in standard regression models. This two-step method is referred as the principal component regression (PCR) (17). PCR is a dimension reduction method, which handles multicollinearity between features and reduces

overfitting of the regression model. Other approaches include the partial least squares (PLS), which uses a dimension reduction technique to investigate the association between two sets of variables (18, 19). Given a vector of predictors, X, and a (potentially multivariate) outcome variable, Y, PLS looks for linear combinations of the components of X that maximize the covariance with Y (or linear combinations of Y if the outcome is multivariate).

Statistical methods have been developed or adapted to suite the characteristics specific to each *-omics* data. For example in epigenetics dataset, each feature has a physical position on the chromosome, so that the features can be ordered in each chromosome. Instead of studying each feature independently, it is thus possible to investigate regions of interest using the hypothesis that neighboring features may share similar information. Statistical methods specific to epigenetics data such as the differentially methylated regions (DMRs) analyses have been developed to that end. The DMRs analysis rationale is to identify regions by combining results from feature-specific analysis for a specific chromosome and using distances between features as weights (20). Other methods such as the fused lasso (FL) regression can be adapted to suite epigenetics data. FL is a generalization of the Lasso, which is well suited when features are naturally ordered (Tibshirani et al., 2005). FL is a multivariable regression method which combines two penalties: (i) the Lasso penalty, which encourages sparsity, i.e. many elements of the estimated vector are encouraged to be set to zero, and (ii) the fused penalty, which encourages sparsity of the difference between two consecutive features, thus introducing smoothness in the parameter vector.

DNA methylation

With hundreds of thousands features measured, epigenetics is the *-omics* set with the highest numbers of variables. It was first introduced by Conrad Waddington in the 1940s. He defined epigenetics as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (21). Several other definitions were then proposed following the new understanding of the mechanisms underlying gene regulation and cell specification (22). In 2008, a new consensus definition of epigenetics term as “stably heritable phenotypes resulting from changes in a chromosome without changes in gene sequence” has been proposed (23), and it is now widely accepted. In other words, epigenetics aims at investigating changes in gene activity not attributable to changes in the DNA sequence. Epigenetics regulates gene transcription, determining where and when a gene is switched on, together with its level of activity. For example, during female embryogenesis, mammalian females randomly inactivate one of their two X-chromosomes via an epigenetic mechanism called X-chromosome inactivation, which causes the transcriptional silencing of one of the two X chromosomes in each female cell

(24). Although every cell in the organism contains the same genetic information, epigenetics can be responsible for different levels of expression of genes in different cells types, i.e. not all genes are expressed simultaneously by all cell types. The study of epigenetic mechanisms encompasses the study of different markers such as chromatin and histone modifications and non-coding RNAs and DNA methylation, which are the most studied epigenetic markers.

DNA methylation is a mechanism of epigenetic regulation that involves the addition of methyl groups (-CH3), most commonly, to the cytosine of a cytosine-guanine (CpG) DNA sequence to form a 5-methylcytosine (5mC) (Figure 1). Even if the role of DNA methylation in gene expression is not fully understood yet (25), it is an important component in numerous cellular processes, including regulation of tissue-specific gene expression, embryonic development, genomic imprinting and preservation of chromosome stability. Moreover, DNA methylation is suspected to play different roles in gene activity based on its genomic location (26). For example methylated CpG sites located in an island region, i.e. region with a high density of CpG sites (Figure 2), are generally associated with gene repression, especially if the island is located in a promoter gene. Methylated CpG sites located in a gene body region, i.e. between the ATG and stop codons, are more likely to be associated with a higher level of gene expression in dividing cells (27). DNA methylation levels at one CpG site are frequently expressed as the percentage of cells that are methylated at that specific site. The Illumina Infinium HumanMethylation450K BeadChip (HM450K) quantifies DNA methylation at more than 450,000 interrogated CpG sites, expressing methylation levels as the ratio of the methylated probe intensity to the overall intensity, which is the sum of the methylated and unmethylated probe intensities (28). In mammals, around 70% to 80% of CpG sites are methylated in somatic cells (29, 30).

Figure 1. Mechanisms of inheritable epigenetics. (31)

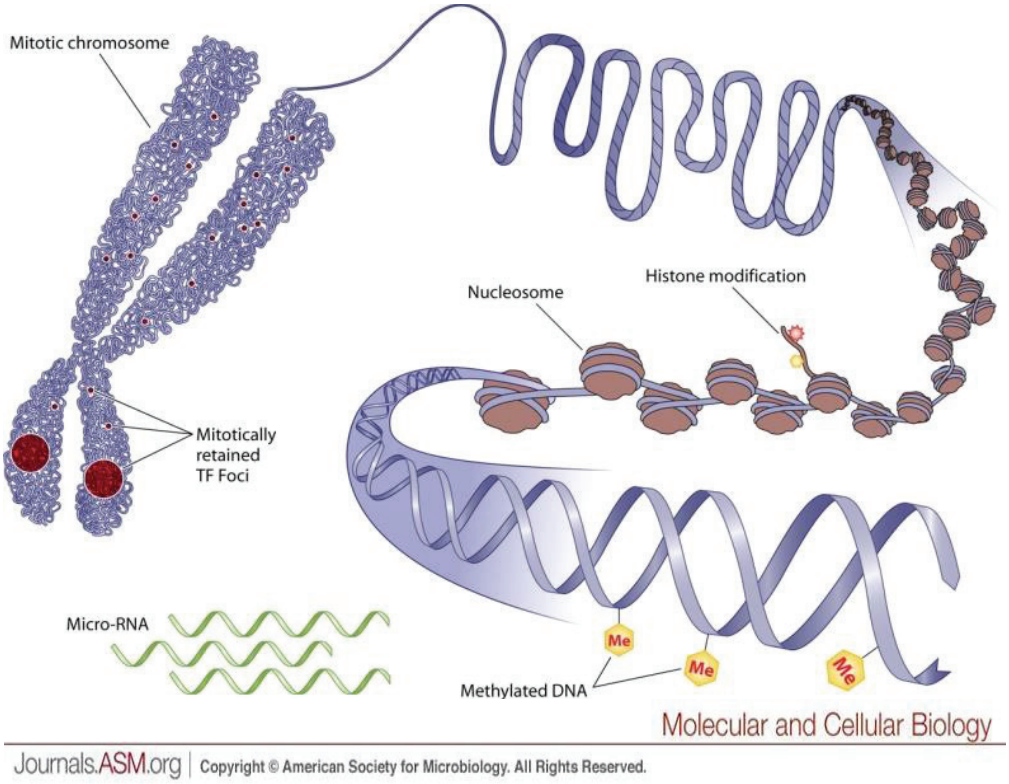
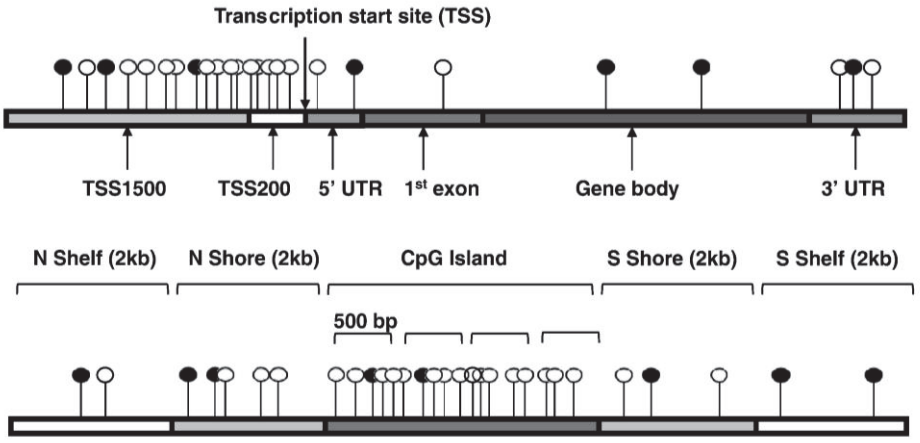


Figure 2. Schematic diagram of gene regions and CpG island regions (28).



Unlike the DNA sequence, which is stable over time, DNA methylation may fluctuate over an individual's lifetime. Embryogenesis and early postnatal life are especially sensitive to DNA methylation changes. Methylation alterations are amplified during these periods, as a consequence of the importance of cell division and somatic maintenance that might affect a high proportion of cells in the development of the organism. During early fetal development, parental methylation profiles or exposures *in utero,* including mother's level of obesity or dietary exposures, are also involved in the embryo methylation changes (32). During embryogenesis abnormal methylation may occur and conduct to abnormal expression or

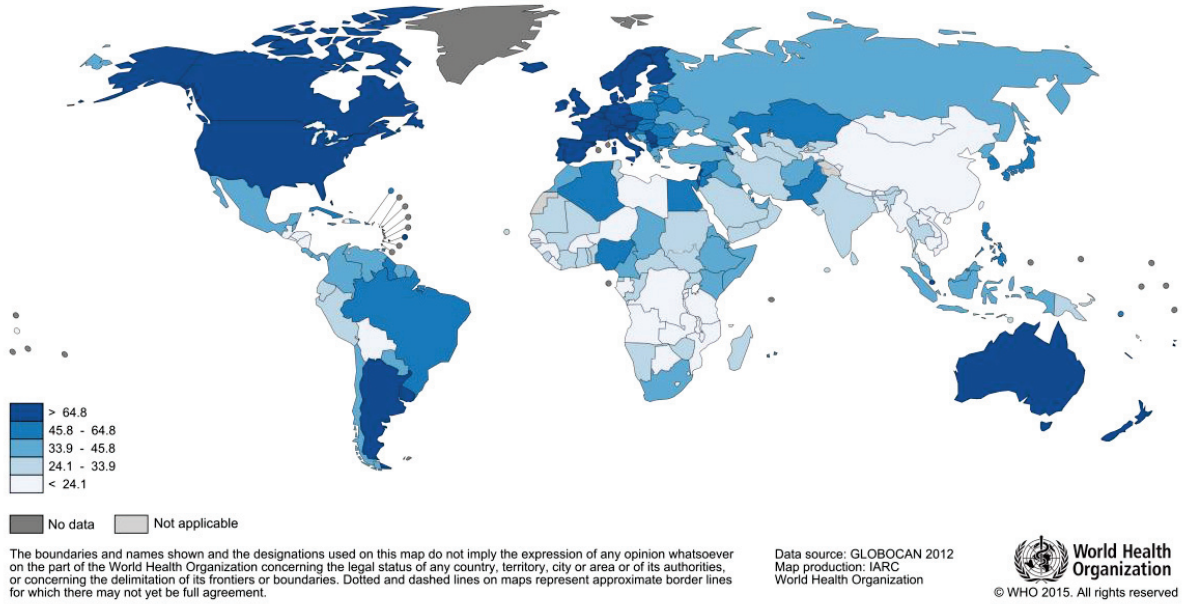
silence of certain genes, which may affect growth and development, and increase the risk of chronic diseases later in life, such as the development of cancer. During life-course, age and specific environmental exposures contribute to changes in DNA methylation, which, in turn, might have long-term effects on development, metabolism and health (33, 34). In this respect, there is increasing evidence supporting an effect of smoking (35, 36), obesity (37, 38) and specific dietary factors (39, 40) on DNA methylation changes.

Due to the important role of DNA methylation in the regulation of many cellular processes, abnormal DNA methylation has been associated with a growing number of human diseases (41). In particular, DNA methylation is suspected to be involved in the development of autoimmune diseases including type I diabetes (42), inflammations associated with cardiovascular disease (43), hypertension (44), respiratory diseases such as asthma and chronic obstructive pulmonary disease (COPD) (45). The role of epigenetic changes in the dysregulation of a wide range of key cellular processes has emerged in many cancer types (46, 47), including breast cancer (48), colorectal cancer (49) and lung cancer (50). Cancer cells are characterized by global hypo-methylation and regional hyper-methylation of CpG islands, which may inactivate fundamental cellular processes such as DNA repair, cell cycle, cell invasion and cell adherence (51). More specifically, DNA hypo-methylation is associated in particular with unusual gene reactivation leading to a potential overexpression of some normally silenced genes such as oncogenes, which might for example increase proliferation of cancerous cells. DNA hyper-methylation is frequently associated with gene repression and genomic instability (through silencing of DNA repair genes) and may result in silencing of important genes, such as tumor-suppressor genes.

Breast cancer

With 1,677,000 newly diagnosed cases in 2012, breast cancer (BC) is the most frequent cancer among women worldwide (52). Before the age of 75 years, 1 in 22 women will be diagnosed with BC. Even if the incidence rates vary nearly four-fold across world regions, BC represents about 25% of all cancers in women (Figure 3). BC is the most frequent cause of cancer death in women in less developed countries and the second cause of cancer death in more developed regions after lung cancer. It is the fifth most common cause of death from cancer overall (522,000 deaths in 2012). Age standardized incidence and mortality rates were respectively 43.3 and 12.9 per 100,000 in 2012. A quarter of BC cases and deaths in the world occurred in Europe where the 5-year relative survival rate ranged between 71% in Latvia and 87% in Finland (53).

Figure 3. Age standardized breast cancer incidence rates in the world in 2012.



Source: http://globocan.iarc.fr

The most common BC type is the invasive breast adenocarcinoma, for which cancer cells start growing in the breast ducts or glands and then spread into the surrounding breast tissues. Different treatment protocols exist for invasive BC, including chemotherapy, hormone or targeted therapies, radiation and surgery. In order to determine the appropriate treatment options, it is important to know the status of the hormone receptor for estrogen (ER), progesterone (PR) and human epidermal growth factor 2 (HER2). Hormone receptor-positive BC cells have either estrogen-positive (ER+) or progesterone-positive (PR+) receptors. These two cancers respond to hormone therapy drugs that lower estrogen levels or block estrogen receptors, preventing the cancer cells from getting the hormones levels they need to further grow. ER+ receptor is expressed in approximately 80% of invasive BC and has a more favorable initial prognosis than ER-. Hormone receptor-negative cancers have neither estrogen nor progesterone receptors and tend to grow faster than hormone receptor-positive cancers. HER2-negative BC (HER2-) have little or no HER2 protein, while this protein is over-expressed in HER2-positive (HER2+) cancers. HER2 protein is involved in the pathway for cell growth and survival. Triple-positive cancers (ER+, PR+ and HER2+) can be treated with hormone drugs, as well as drugs that target HER2 whereas chemotherapy is needed for triple-negative cancers (ER-, PR- and HER2-) as hormone therapy is not helpful in treating these cancers because of the absence of hormonal receptors and low levels of HER2. Hormone receptors and HER2 expression inform on the choice of the treatment once invasive cancer has been diagnosed as part of a second prevention scheme.

BC is a multifactorial disease with several well identified risk factors (54-56), including hormonal and reproductive factors such as age at menarche and menopause, parity, breastfeeding, use of oral contraceptives and hormonal menopausal therapy. Known non-modifiable factors of BC include height, age, ionizing radiation and genetic factors including family history. BC has very few well established modifiable risk factors. Lifestyle and environmental factors, such as alcohol consumption (57), obesity and physical activity (58), are suspected to contribute to BC risk (59) but results are still scarce and inconsistent. The identification of modifiable risk factors is a current research topic aiming to strengthen primary prevention.

Numerous studies suggested an association between specific nutrients and BC, particularly fatty acids, carbohydrates, vitamins B, vitamin D, carotenoids, phytoestrogens, and dietary fibers (60). The WCRF review panel concluded that the epidemiological data for an association between folate and BC risk was too limited to allow for conclusions (59). However recent epidemiological studies supported evidence of a protective effect of folate on BC risk (61), even if the mechanisms through which folate operates are yet not fully understood. Specific subtypes of fatty acids have been also posited to affect BC risk. In particular, high levels of palmitoleic acid, used as a biomarker of endogenous lipogenesis, have been associated with an increased risk of BC (62), so were an increased levels of trans-fatty acids for ER-negative tumors.

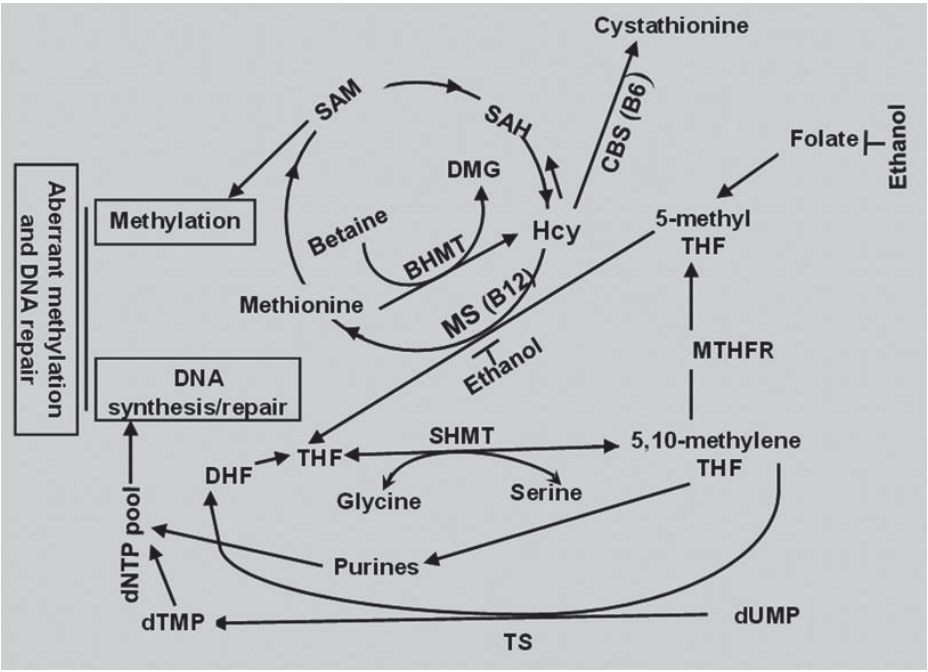
The role of DNA methylation in breast cancer occurrence

Dietary factors may change epigenetics profiles, which in turn may alter the susceptibility to BC. Based on a literature review, the potential role of specific dietary components, including micronutrients such as folate, macronutrients such as alcohol, and soya intake, in modifying BC risk via epigenetic mechanisms has been reviewed recently (63). In light of epidemiological, animal and clinical studies, the role of specific dietary factors to modulate BC risk were discussed, together with candidate underlying mechanisms related to the interaction of diet and the epigenome. Understanding the interplay between nutrition and epigenetics is particularly important as many nutrients have been described to have a potential impact on the mammary gland and its tumorigenesis (64). Specific nutrients may be capable of inducing permanent epigenetic modifications, such as gene repression.

B-vitamins, particularly folate, are essential components of the one-carbon metabolism (OCM). The OCM is a network of interrelated biochemical reactions involved in the donation of methyl groups from nutrients to DNA methylation reactions in the cells, including the methylation of DNA, RNA and proteins (Figure 4). Modifications in OCM can significantly

impact gene expression via epigenetic mechanisms and thereby cellular function (65). Folate is the major source of methyl groups from food. A low folate intake results in a low methyl supply which may lead to a global DNA hypo-methylation caused by high homocysteine concentrations and low methionine regeneration (66). This may increase the susceptibility of genes to mutations or alter gene expression. Inadequate folate level may also result in abnormal DNA synthesis due to a reduced availability of S-adenosyl methionine (SAM) (67) and disrupted DNA repair and hence may influence cancer risk, including BC (68). Indeed, several epidemiological studies suggested a protective role of folate and related B vitamins on BC (69, 70). Folate has been inversely associated with BC risk, possibly reflecting a role of folate to modulate the expression of gene that regulates tumor development and progression.

Figure 4. One-carbon metabolism pathway (67).



CBS, cystathionine b-synthase; DHF, dihydrofolate; dNTP, deoxyribonucleotide triphosphate; DMG, dimethylglycine; dTMP, deoxythymidine monophosphate; dUMP, deoxyuridine monophosphate; Hcy, homocysteine; MS, methionine synthase; MTHFR, methylene THF reductase; 5-methylTHF, SAH, S-adenosyl homocysteine; SHMT, serine hydroxymethyltransferase.

Beside folate, alcohol intake has also been shown to influence epigenetic profiles (71). Ethanol metabolism generates toxins that may reduce folate absorption, mainly by increasing renal excretion of folate and inhibiting methionine synthase, thus leading to OCM dysfunction (71, 72). The antagonist effect of alcohol on folate could increase the need of folate intake, thus indirectly increase BC risk. Recent epidemiological evidences indicated that high alcohol

consumption was positively associated with BC (57, 73), particularly for low levels of folate intake (74).

Thesis objectives

The overall objective of the thesis was to investigate the variability of epigenetics data, possibly separating out variation attributable to technical processing of samples from biological variation, i.e. due to epidemiological factors that are involved in the etiology of breast cancer, including age, smoking, dietary folate, alcohol and fatty acid profiles. The work entailed the use of novel statistical methods to quantify, correct and exploit the variability in methylation level data. Existing methodology was adapted to suite the analysis of large dimension data to progressively acquired novel information of important features of epigenetics data.

DNA methylation measurements may be affected by systematic and random variation due to the processing of samples. The work focused primarily on statistical methodology aiming at identifying sources of random and systematic variability in methylation levels, either introduced by the technical treatment of samples after collection from study participants, or during the acquisition phase by allocation of samples into chips within laboratory batches. The PC-PR2 method was adapted to the analysis of epigenetics data, and that was possibly by exploiting a desirable property of PCA, which is invariant to transposition of the design matrix. The method lent itself as a very handy way to handle very cumbersome data in terms of size of features to process simultaneously. Once the sources of variation were identified and quantified, the thesis focused on the evaluation of the performance of the most popular methods to remove unwanted variability. In order to evaluate the performance of different normalization methods, three different techniques, i.e. ComBat, SVA and a method based on the computation of residuals were compared in terms of their ability to remove unwanted variation. For this purpose, the association between smoking status and DNA methylation within the CHARGE Consortium was used as an application. This work was described in an article that was published in Clinical Epigenetics (F. Perrier, 1st author).

Once the evaluation phase was completed, the work focused on the estimation of the relationship between dietary factors related to BC and methylation levels, complementing standard statistical analysis with more advanced statistical techniques for the identification of specific features of epigenetics data. This objective was subdivided into three parts. First, the association between plasma concentrations of folate and vitamin B12 and BC risk was assessed within a nested case-control study in the EPIC cohort. This study was published on the International Journal of Cancer (F. Perrier, 4th author) by Dr. Marco Matejcic, a post-doctoral epidemiologist. For this study, I participated to the development of the statistical

methodology used in the analysis. Second, the relationship between dietary folate and alcohol intake with DNA methylation patterns was investigated using three statistical approaches: the site-specific analysis, the DMRs analysis and the FL regression. The manuscript of this study has been recently submitted in Clinical Epigenetics (F. Perrier, 1st author). Third, the relationship between average methylation level and BC risk overall and in specific regions of the epigenome reflecting the physical location of CpG sites in relation to CpG islands, was explored using conditional logistic regression models. This work was part of a study coordinated by the IARC Epigenetics Group, in which I was involved for the development and implementation of statistical analyses. An article from Dr. Srikant Ambatipudi was published in the European Journal of Cancer (F. Perrier, 3rd author).

EPIC study

Data analyzed in this thesis were derived from the European Prospective Investigation into Cancer and nutrition (EPIC) cohort. The EPIC study is a multicentre study that recruited over 521,000 study participants, between 1992 and 2000 in 23 regional or national centres in 10 European countries (Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and United Kingdom) (75). The main aim of the EPIC study is to investigate the etiology of cancers at many sites in relation to diet and lifestyle factors using prospective centre-specific data. Information was collected at recruitment via a lifestyle and health factors questionnaire and a validated centre- or country-specific dietary questionnaire to capture local dietary habits. Anthropometric measurements were performed for all participants and blood samples were taken at recruitment for approximately 400,000 EPIC participants. A 24-hours dietary recall was implemented in a total of 36,900 participants from each centre in order to calibrate dietary measurements. From the recruitment of study participants from 1992 to 1999 until the end of the follow-up in 2009, 47,000 EPIC participants were diagnosed with cancer (Table 1).

Table 1. Number of incident cancers and deaths in EPIC in 2010

Country	*N*	Person-Years	No. of incident cancers	No. of incident deaths
France	74 524	1 103 492	7313	4038
Italy	47 745	582 716	3862	1708
Spain	41 438	562 044	2887	1972
United Kingdom	87 887	1 110 137	8301	9587
Netherlands	40 011	509 852	3170	2386
Greece	28 561	266 099	1137	2146
Germany	53 088	595 857	4443	2836
Sweden	53 823	742 397	6806	5780
Denmark	57 053	664 510	7249	5549
Norway	37 200	406 473	2357	975
Total	**521 330**	**6 543 577**	**47 525**	**36 977**

Source : http://epic.iarc.fr/about/cohortdescription.php

Among the 367,903 women recruited in EPIC, 19,583 participants had prevalent cancers at recruitment (except non-melanoma skin cancer) and 2,892 women were lost during follow-up. First malignant primary BC occurred for 10,713 women of the EPIC cohort during the follow-up time. A nested case-control study was designed among women who completed dietary and lifestyle questionnaires and provided blood samples at recruitment (baseline), which included 3,858 invasive BC cases. Each case was matched to a randomly selected control among cancer-free women by recruitment centre and the following baseline variables: age, menopausal status, fasting status, current use of oral contraceptive pill or hormone replacement therapy and time of blood collection (76).

Within the BC nested case-control study, a subsample of 960 women (480 cases and 480 matched controls) from Germany, Greece, Italy, Netherlands, Spain and United Kingdom was selected for the DNA methylation analysis (77).

PART I: Normalization approaches to correct for systematic sources of variation in DNA methylation measures

Context

DNA methylation is altered by many factors including age (34) and environmental factors (78) such as smoking (35, 36) and alcohol consumption (67, 71). But systematic and random variation introduced by the technical processing of biospecimens might also affected methylation measures. This may compromise the accuracy of the measurement process and contribute to bias the estimate of the association under investigation. It includes in particular variability attributed to batch (a group of 96 samples processed at the same time), chip position within batches (8 chips per batch) and the position of the samples within the chip (12 samples per chip allocated into 2 columns and 6 rows) (4). The quantification of the contribution of the sources of systematic and random variation is challenging in datasets characterized by hundreds of thousands of features.

Objectives

- To identify and quantify the contribution of systematic and random sources of variation in methylation measurements.
- To evaluate the performance of three normalization techniques accounting for unwanted variability in methylation measurement using the association between smoking and DNA methylation levels.

Approach

Illumina Infinium HumanMethylation450K was used to acquire methylation levels in over 421,000 CpG sites for 902 buffy coat samples from study participants of a case-control study on BC nested within the EPIC cohort. Smoking status was categorized into never vs ever smokers based on lifestyle questionnaires.

In this study, the principal component partial R-square (PC-PR2) analysis (3), a method previously developed for the analysis of metabolomics data was introduced to evaluate the performance of normalization techniques to correct for unwanted variation. The PC-PR2 method was used to identify and quantify the contribution of laboratory factors and other characteristics of the samples variability, before and after each of the normalization technique. Three normalization techniques, namely ComBat (9), surrogate variables analysis (SVA) (12, 13) and a residuals approach based on the computation of residuals from regression model were performed on raw β-values and M-values. Sites-specific analyses

evaluating the association between smoking status and DNA methylation levels after application of each of the three normalization methods were performed. Results were compared with findings from the CHARGE consortium, a large meta-analysis combining pooled data from 16 cohorts and including about 16,000 samples (36).

Main findings

For β-values, a sizeable proportion of variability attributable to variables expressing batch and row sample position within chip was identified, with values of the partial R2 statistics equal to 9.5% and 11.4% of total variation, respectively. After application of ComBat or the residuals' methods, the contribution was 1.3% and 0.2%, respectively. The SVA technique resulted in a reduced variability attributable to batch (1.3%) and row sample position (0.6%), and in a reduced variability attributable to chip within a batch (0.9%). Similar results were obtained for M-values.

Using standard adjustment and FDR correction of p-values, i.e. models using the raw methylation values and adjusted for batch and row sample position, smoking status was significant associated with changes of methylation levels in 444 sites, 80% of which were overlapping results from the CHARGE consortium. After ComBat and the residuals' normalizations, a larger number of significant sites (k = 600 and k = 427, respectively) were associated with smoking status than after SVA correction (k = 96). However, almost all the significant sites after SVA were overlapping results from the CHARGE consortium (96%) compare to ComBat and the residuals methods, 69% and 85% respectively. Similar results were obtained for M-values with a higher percentage of overlapping sites.

Conclusion

Our findings suggested that laboratory factors such as the position of the sample within the chip and the position of the chip within batches can add unwanted variability to DNA methylation in addition to the variability introduced by the batch. In an analysis of EPIC data, the PC-PR2 method lent itself as a very useful tool to explore the contribution to total variability of an *a priori* list of laboratory factors and sample characteristics. This step turned out to be essential to evaluate the performance of routinely used normalization methods, such as the regression-based residuals, ComBat and SVA, and to further appreciate the extent of these corrections. SVA produced more conservative findings than ComBat and the residuals' methods in the association between smoking and DNA methylation. These steps should be part of the pre-processing analysis of any *-omics* data.

Published article: Identifying and correcting epigenetics measurements for systematic sources of variation.

Perrier *et al. Clinical Epigenetics* (2018) 10:38
https://doi.org/10.1186/s13148-018-0471-6

Clinical Epigenetics

METHODOLOGY Open Access

Identifying and correcting epigenetics measurements for systematic sources of variation

Flavie Perrier[1], Alexei Novoloaca[2], Srikant Ambatipudi[2], Laura Baglietto[3], Akram Ghantous[2], Vittorio Perduca[4], Myrto Barrdahl[5], Sophia Harlid[6], Ken K. Ong[7], Alexia Cardona[7], Silvia Polidoro[8], Therese Haugdahl Nøst[9], Kim Overvad[10,11], Hanane Omichessan[12,13], Martijn Dollé[14], Christina Bamia[15,16], José Marìa Huerta[17,18], Paolo Vineis[19], Zdenko Herceg[2], Isabelle Romieu[20] and Pietro Ferrari[1*]

Abstract

Background: Methylation measures quantified by microarray techniques can be affected by systematic variation due to the technical processing of samples, which may compromise the accuracy of the measurement process and contribute to bias the estimate of the association under investigation. The quantification of the contribution of the systematic source of variation is challenging in datasets characterized by hundreds of thousands of features. In this study, we introduce a method previously developed for the analysis of metabolomics data to evaluate the performance of existing normalizing techniques to correct for unwanted variation. Illumina Infinium HumanMethylation450K was used to acquire methylation levels in over 421,000 CpG sites for 902 study participants of a case-control study on breast cancer nested within the EPIC cohort. The principal component partial R-square (PC-PR2) analysis was used to identify and quantify the variability attributable to potential systematic sources of variation. Three correcting techniques, namely ComBat, surrogate variables analysis (SVA) and a linear regression model to compute residuals were applied. The impact of each correcting method on the association between smoking status and DNA methylation levels was evaluated, and results were compared with findings from a large meta-analysis.

Results: A sizeable proportion of systematic variability due to variables expressing 'batch' and 'sample position' within 'chip' was identified, with values of the partial R^2 statistics equal to 9.5 and 11.4% of total variation, respectively. After application of ComBat or the residuals' methods, the contribution was 1.3 and 0.2%, respectively. The SVA technique resulted in a reduced variability due to 'batch' (1.3%) and 'sample position' (0.6%), and in a diminished variability attributable to 'chip' within a batch (0.9%). After ComBat or the residuals' corrections, a larger number of significant sites ($k = 600$ and $k = 427$, respectively) were associated to smoking status than the SVA correction ($k = 96$).

Conclusions: The three correction methods removed systematic variation in DNA methylation data, as assessed by the PC-PR2, which lent itself as a useful tool to explore variability in large dimension data. SVA produced more conservative findings than ComBat in the association between smoking and DNA methylation.

Keywords: Epigenetics, PC-PR2, Normalization, Methylation, Smoking status

* Correspondence: ferrarip@iarc.fr
[1]Nutritional Methodology and Biostatistics Group, International Agency for Research on Cancer (IARC), World Health Organization, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France
Full list of author information is available at the end of the article

Background

Epigenetics aims at investigating changes in gene activity not attributable to changes in the DNA sequence [1]. An increasing number of studies analysed epigenetics in relation to modifiable environmental exposures of epidemiologic interest, such as smoking [2–4], alcohol consumption [5], maternal plasma folate [6] and other vitamin involved in the one carbon metabolism pathway [7], as well as the role of epigenetic profiles on the risk of developing chronic diseases, including cancer [8]. DNA methylation is a mechanism of epigenetic regulation that involves the addition of methyl groups (–CH3) to the cytosine of a cytosine-guanine DNA sequence. DNA methylation level at one CpG site is frequently expressed as the percentage of cells that are methylated at that specific site. The Illumina Infinium HumanMethylation450K BeadChip (HM450K) quantifies DNA methylation at more than 450,000 interrogated CpG sites, expressing methylation level as the ratio of the methylated probe intensity to the overall intensity, which is the sum of the methylated and unmethylated probe intensities [9].

Methylation levels are influenced by many factors including aging [10] and environmental exposure [11, 12], but might also be affected by systematic variation due to the processing of the biospecimens, e.g. variability attributed to batch (a sub-group of samples processed at the same time, 96 samples per batch in the HM450K), chip position within batches (8 chips per batch in the HM450K) and the position of the samples within the chip [13]. Methods of correcting for the sources of methylation variability include ComBat, based on an empirical Bayes method [14] and the surrogate variables analysis (SVA) [15, 16]. An alternative method consists in the computation of residuals from a beta regression, where methylation levels were regressed on the major sources of methylation variability.

The large dimension of new generation methylation arrays makes it difficult to quantify the amount of variability attributable to systematic sources of variation. The principal component partial R-square (PC-PR2) method was developed to quantify the contribution of sources of variation defined a priori in large dimensional data [17].

Smoking exposure has been analysed in many studies [2–4], which offers a large comparative pool of results. Smoking has also been shown to have a major impact on the epigenome and hence provides a large number of significant CpGs to analyse. For these reasons, in this work, we have chosen to evaluate the performance of ComBat, SVA and the residuals' method to correct for potential systematic variability in methylation measurements, in the association between smoking and DNA methylation levels from DNA samples of subjects of a nested case-control study on breast cancer conducted within the European Prospective Investigation into Cancer and nutrition (EPIC) study. The PC-PR2 method was used to quantify the extent of total epigenetics variability before and after applying each correcting method.

Methods

Study population

The EPIC study [18, 19] is a multicentre study that recruited over 521,000 study participants, between 1992 and 2000 in 23 regional or national centres in 10 European countries (Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and the UK). Among the 367,903 women recruited in EPIC, we excluded 19,583 participants with prevalent cancers at recruitment (except non-melanoma skin cancer) and 2892 women that were lost during follow-up. Malignant primary breast cancer (BC) occurred for 10,713 of them from 1992 to 2010. A nested case-control study was designed among women who completed dietary and lifestyle questionnaires and provided blood samples at recruitment (baseline), which included 3858 invasive BC cases. Each case was matched to a randomly selected control among cancer-free women by recruitment centre and the following baseline variables: age, menopausal status, fasting status, current use of oral contraceptive pill or hormone replacement therapy and time of blood collection [20].

Genome-wide DNA profiling assessment

Genome-wide DNA-methylation profiles in buffy coat samples was quantified using the Illumina Infinium HumanMethylation450K (HM450K) BeadChip assay [9] in 960 biospecimens of women included in the BC nested case-control study [21]. The 480 cases were selected based on estrogen receptor status and by selecting equal proportions of subjects with above or below median level of dietary folate. Matched controls were the same than those selected for the whole study. A total of 20 biospecimens with replicates were used to compare technical inter- and intra-assay batch effects and then excluded from the main analysis. We also excluded 19 matched pairs where at least one of the two samples had a low-quality bisulfite conversion efficiency (intensity signal < 4000) or which did not pass all the Illumina GenomeStudio quality control steps, which were based on built-in control probes for staining, hybridization, extension, and specificity [22]. A total of 451 completed matched pairs ($n = 902$) were retained for the main statistical analyses. In any given sample, probes with detection p value higher than 0.05 were assigned 'missing' status. After the exclusion of 14,548 cross-reactive probes, 47,963 probes overlapping known SNPs with minor allele frequency (MAF) of ≥ 5% in the overall population (European ancestry) [23] and 1483 low-

quality probes (missing in more than 5% of the samples), 421,583 probes were included in the statistical analyses.

For each probe, β value was calculated as the ratio of methylated intensity and the overall intensity, defined as the sum of methylated and unmethylated intensities. The following preliminary adjustment steps were applied to the β values: (i) color bias normalization using smooth quantile normalization to correct for the two color channels; (ii) quantile normalization [24]; (iii) type I and type II bias correction using the beta-mixture quantile normalization (BMIQ) [25]. Then, M values, defined as $M_{\text{values}} = \log_2\left(\frac{\beta_{\text{values}}}{1-\beta_{\text{values}}}\right)$, were computed [26]. In this work, the β and M values obtained after the preliminary normalization steps were referred to as the raw β and M values.

The amount of white blood cell counts (T cells ($CD8^+T$ and $CD4^+T$), natural killer (NK) cells, B cells, monocytes and granulocytes) was quantified using Houseman's estimation method [27]. The percentage of granulocytes was not included in this analysis as it is collinear with the five other white blood cell counts: the total of the percentages of the six leukocyte subtype counts is 1.

For the DNA methylation measurements with the HM450K BeadChip, samples were aliquoted into 10 batches; each batch was made of 8 chips, and each chip contained 12 samples (located in 2 columns of 6 rows). Chip position represented the position of the chips within a batch, as illustrated in Fig. 1a, and sample position represented the position of the samples within a chip, as in Fig. 1b.

Lifestyle exposures

Data on lifestyle exposures were collected at recruitment through country- or centre-specific dietary and lifestyle questionnaires [18]. Smoking status was categorized into ever (former/current) and never smokers and was not associated to any of the technical covariates.

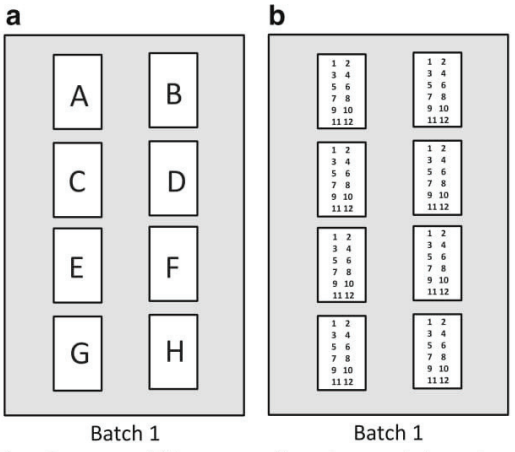


Fig. 1 Description of laboratory variables. **a** Position of chips within batches, each batch was made of 8 chips. **b** Sample position within chips, each chip contains 12 samples

Statistical analyses

In order to inspect the variability of DNA methylation levels, we first visually inspected, via box plots, global DNA methylation levels by batch, chip and sample positions. The principal component partial R-square (PC-PR2) method was used to quantify the contribution of laboratory factors and other characteristics of the samples to the between-sample variability observed [17]. First, principal component analysis (PCA) was carried out, by the PC-PR2, on the matrix X of epigenetics data of dimension $n \times p$ ($n = 902$: number of study samples and $p = 421{,}583$: number of probes). In PCA, eigenvalues and eigenvectors are usually obtained from the matrix $X'X$ of dimension $p \times p$. In this case, and in general with *-omics* data, p is very large ($p \gg n$), and the decomposition of $X'X$ can be cumbersome. A particularly appealing procedure consists in extracting eigenvalues and eigenvectors from the matrix XX', of dimension $n \times n$ [28], which is way easier to handle, being n much smaller than p. Once eigenvalues were extracted, the q first components explained an amount of total variability in X greater than a given threshold, i.e. 80% in this study. Then, each of the q first PCA score components was, in turn, linearly regressed on a list of independent covariates (Z), comprising of laboratory factors and characteristics of the samples. Values of the partial R^2 statistics were assessed for each Z covariate, separately in each component-specific model [29]. An overall partial R^2 was computed for each Z covariate with a weighted average of their component-specific partial R^2 using the corresponding q eigenvalues as weights, conditional to all other covariates in the model. The covariates that we have entered into the regression include batch, chip position, row sample position, recruitment centre, proportions of leukocyte subtypes ($CD8^+T$, $CD4^+T$, NK, B cells and monocytes), alcohol consumption (g/day), age (year), BMI (kg/m^2), menopausal status (post- vs. pre-menopause), smoking (ever vs. never smokers), BC status (case or control) and dietary folate intake (μg/day).

Removing unwanted variation

To remove the two most important sources of variation identified with the PC-PR2 from DNA methylation levels, three different correcting techniques were applied to raw β and M values: residuals, ComBat and SVA. The ComBat method [14] is a procedure based on an empirical Bayes approach that can correct only for one covariate at the time. Given the presence of multiple sources of variation, we have applied two parametric ComBat in multiple sequential steps: ComBat was first applied to remove batch variability, and then a second ComBat step

was run to remove variability due to row sample position. Methylation β values that after the application of ComBat were lower than 0 or larger than 1 were set to 0 and 1 respectively. The surrogate variables analysis (SVA) is a method developed to remove pre-identified sources of variability but also non-known sources of variability, i.e. variability which is not specified in the SVA model, using surrogate variables [15, 16]. Once surrogate variables were assessed by SVA, residuals from a regression modeling methylation level according to the surrogate variables were computed to remove the unwanted variation.

As the β values are continuous in the [0,1] interval, the calculation of the residuals for the residuals' method and SVA method were based on beta regression. To be comparable to the ComBat and raw (i.e. uncorrected) data, residuals computed with the residuals' and the SVA methods needed to be rescaled as follows:

$$res_{\text{scaled},j} = \frac{\text{res}_{\text{raw},j} - \min(\text{res}_{\text{raw},j})}{\max(\text{res}_{\text{raw},j}) - \min(\text{res}_{\text{raw},j})} \left(\max(\text{raw}_j) - \min(\text{raw}_j) \right) + \max(\text{raw}_j)$$

where $j = 1\ldots421{,}583$, raw_j represents the raw β values measured in site j and $\text{res}_{\text{raw},\,j}$ the residuals computed for site j before transformation.

In order to check the efficacy of the three correcting techniques, a second PC-PR2 analysis was used to quantify the contribution of each laboratory factor to total variability, after each of the normalization methods.

Same approach was used for M values using a linear regression instead of beta regression to compute residuals from the residuals' and the SVA methods.

In order to compare sample individual values before and after correction, raw and corrected β and M values of the probe cg00000029 were visually inspected. In this site, in addition to the three tested methods, a second residuals' method was also computed using random effects instead of fixed affects to remove unwanted variation, from a beta or linear mixed regression, respectively for β and M values.

CpG site-specific models

The association between smoking status and each of the 421,583 CpG sites was carried out before and after application of each normalization method. Beta regression models were used for β values and linear regression models for M values, with adjustment for chip position, recruitment centre, percentages of five leukocyte subtypes, age at recruitment, menopausal status and BC status. The standard adjustment models, i.e. models using the raw methylation values, were also adjusted for batch and row sample position. In order to compare the epigenome-wide distribution of p values with the expected null distribution of p values, the inflation factor λ was computed and the quantile-quantile (QQ) plots were generated. The inflation factor was defined as the ratio of the median of the observed $\log_{10}$ transformed p values and the median of the expected $\log_{10}$ transformed p values. False discovery rate (FDR) was used to control for multiple testing. In order to compare the performance of the different correction methods with a nominal reference, the list of k significant CpG sites (q values < 0.05) associated with smoking was compared to the results of a large meta-analysis carried out in the CHARGE consortium, a recent large meta-analysis on the link between the epigenetic signature of cigarette smoking that pooled data from 16 studies, and included about 16,000 individuals [4]. In CHARGE, smoking status was statistically significantly associated with DNA methylation level (β values) in 18,760 sites, after FDR correction of p values.

In order to compare the performance of the correction methods, the relative sensitivity and specificity of each correcting method were computed. We considered the CpG sites significantly associated to smoking in the CHARGE consortium as the true positives, i.e. an arbitrary gold standard, given that this is a well-powered reference study and the largest to date.

Preprocessing steps and statistical analysis were carried out using the R software (https://www.r-project.org/) and Bioconductor packages [30], including 'lumi' and 'wateRmelon' for the adjustment step, 'sva' [31] for ComBat and SVA corrections, and 'betareg' for beta regression models. The PC-PR2 method was computed using the R code available in Fages et al.'s supplementary material [17].

Results

DNA measurements of the first and the last batches were conducted roughly 3 months apart. DNA measurement of two consecutive batches varied from 3 to 14 days. Box plots of global methylation (i.e. mean of methylation levels in all the CpG sites) showed a random variation of global methylation levels between batches, as reported in Fig. 2a for β values. Global methylation between chip positions did not present large variation (Fig. 2b). Sample position within the chip systematically influenced global methylation, with levels by rows, showing a progressive constant increase in methylation, a feature not observed by column, as displayed in Fig. 2c. The impact of row sample position on global methylation was even stronger when batches were evaluated separately (Fig. 2d). Global methylation computed with M values gave similar results (Additional file 1: Figure S1).

Tables 1 and 2 show the results of PC-PR2 to quantify the amount of total variability of DNA methylation explained respectively by laboratory factors and characteristics of the samples (recruitment centre, the five

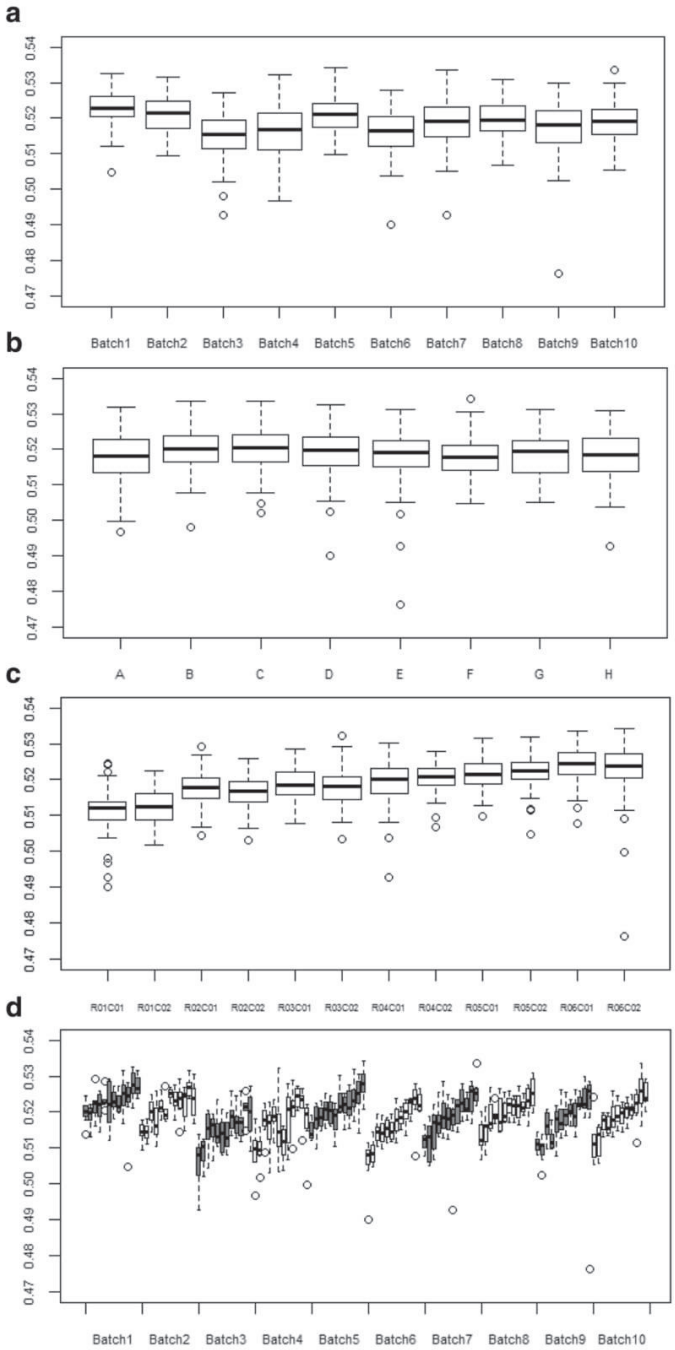


Fig. 2 Box plots of global methylation (β values) according to laboratory factors. **a** Batch. **b** Chip position within batches. **c** Sample position within chips. **d** Batches and sample position within chips

Table 1 Values of weighted partial R^2 (%) from PC-PR2 analysis indicating the proportion of variability of methylation levels, before and after normalization step, explained by a specific set of laboratory factors

Values	Methods[a]	Row sample position	Batch	Chip position	Total[b]
β values	Raw	11.4	9.5	6.5	30.4
	Residuals	0.2	1.3	5.9	17.9
	ComBat	0.2	1.3	6.0	17.1
	SVA	0.6	1.3	0.9	6.5
M values	Raw	12.3	9.7	6.8	30.7
	Residuals	0.2	1.2	5.8	16.5
	ComBat	0.2	1.3	6.2	17.0
	SVA	0.4	0.7	0.8	5.3

[a]Residuals, COMBAT and SVA methods used to correct effect due to batch and row sample position (within the chips)
[b]Total variability explained by laboratory factors and characteristics of the samples (recruitment centre, the five percentages of leukocyte subtypes, alcohol consumption, age and BMI, menopausal status, smoking, BC status and dietary folate)

percentages of leukocyte subtypes, alcohol intake, age, BMI, menopausal status, smoking, breast cancer status and diet folate intake), for raw β and *M* values. Findings were similar for raw β and *M* values; the largest contribution to the overall variability came from row sample position and batch explaining, respectively, 11.4 and 9.5% (β values), and 12.3 and 9.7% (*M* values) of overall methylation variation. Chip position contributed to 6.5 and 6.8%, for raw β and *M* values respectively. The percentages of leukocyte subtypes and centre explained most of the variation of DNA methylation due to sample characteristics for raw β and *M* values. Each of the remaining tested other sample characteristics explained less than 0.5% of total variation.

Table 2 Values of weighted partial R^2 (%) from PC-PR2 analysis indicating the proportion of variability of raw methylation levels explained by a specific set of covariates

Characteristics of samples	β values	*M* values
Recruitment centre	3.0	2.9
Percentages of leukocyte subtypes		
CD4T	3.2	3.2
CD8T	3.7	3.1
Natural killers	5.2	4.7
B cells	1.7	1.1
Monocytes	0.4	0.4
Alcohol intake at recruitment	0.2	0.1
Age at recruitment	0.4	0.4
BMI at recruitment	0.1	0.1
Menopausal status	0.2	0.2
Smoking status	0.1	0.2
Breast cancer status	0.1	0.1
Dietary folate	0.1	0.1

Removing unwanted variation

All the three correcting methods decreased the contribution of row position and batch to similar neglectable levels, whereas only SVA appeared to reduce the contribution to variability due to chip position (Table 1). The amount of variability explained by laboratory factors and sample characteristics for raw β values decreased from 30.4 to 17.9% and 17.1% using, respectively, the residuals' method and ComBat, and to 6.5% after SVA. The PC-PR2 approach applied on *M* values estimated values of partial R^2 for laboratory factors and sample characteristics similar to those of β values.

Corrected methylation values of the probe cg00000029 were very similar using ComBat or the residuals' methods for β values and *M* values (Fig. 3). SVA corrected values were the corrected values most different from the raw values. Using the residuals' method with fixed or random effects for batch and row sample position gave similar results.

CpG site-specific models

The frequency *k* of sites associated with smoking status is shown in Table 3, consistently for β and *M* values. For β values adjusted by batch and row sample position (standard adjustment), smoking status was significantly associated to methylation levels in 444 sites. The number of CpG sites significantly associated with smoking status was equal to 427 for the residuals' method, 600 for ComBat and 96 for SVA after correction. According to the inflation factors and QQ plots, there was no evidence of inflation for any methods (Additional file 2: Figure S2).

These frequencies were compared to the list of 18,760 sites identified in the CHARGE meta-analysis (Joehanes et al. [4]). A total of 77 sites overlapped across the standard adjustment and the three correcting methods in this study and the sites identified in the consortium, as shown in the Venn diagram for β values in Fig. 4a. In addition to these sites, the standard adjustment, the residuals' method and the ComBat method shared a list of 249 significant sites with CHARGE. The ComBat method resulted in the largest frequency of sites overlapping with results in CHARGE ($k = 411$), but also in the largest percentage of sites not observed in CHARGE (31%). In contrast, SVA identified the lowest number of significant sites ($k = 96$) but the vast majority of them (92%) were also identified in CHARGE.

As for *M* values, 322 sites were associated to smoking using the standard adjustment, $k = 332$ after the residuals' method, $k = 387$ using ComBat, $k = 144$ after SVA correction. A total of 111 sites overlapped all the methods and CHARGE, as shown in Fig. 4b. SVA was

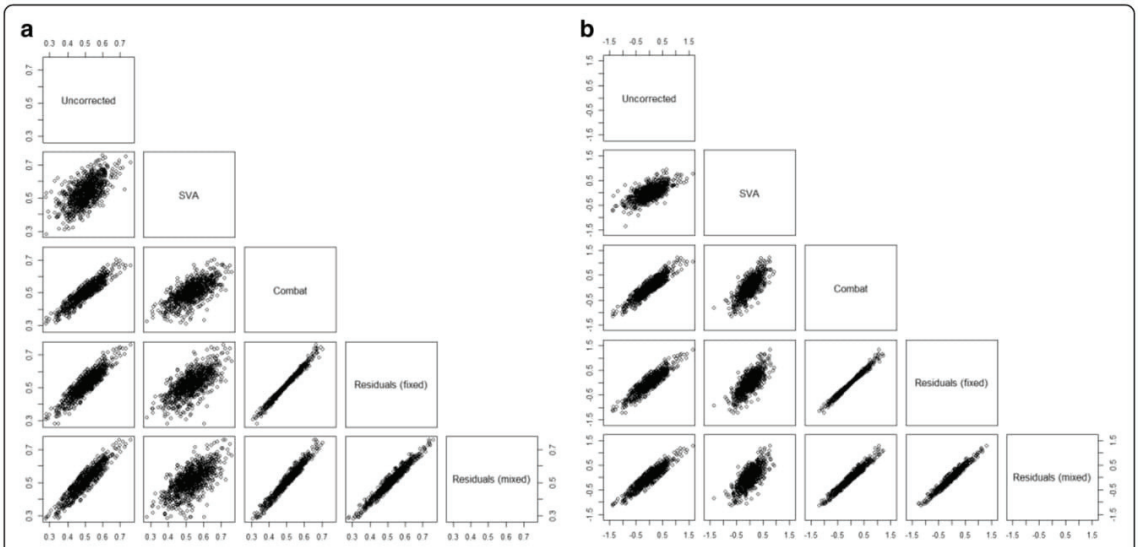


Fig. 3 DNA methylation levels of the CpG site cg00000029 before and after normalization step. **a** β values. **b** *M* values

the method leading to the lowest number of significant sites, but also to the largest percentage of sites also identified by CHARGE (93%). This percentage ranged between 85 and 90% for all the other methods. According to the inflation factors and QQ plots, there was no evidence of inflation for any methods for *M* values (Additional file 3: Figure S3). SVA showed the least inflation in both β values and *M* values.

Sensitivity was similar for the standard adjustment, the residuals' method and the ComBat method with a value about 0.020 for β values and over 0.015 for *M* values (Table 3). SVA sensitivity was four times less for β values and twice less for *M* values. SVA was the most specific method with 1-specificity equals to 0.2×10^{-4} for β values and *M* values whereas ComBat was the least specific with 1-specificity equals to 4.7×10^{-4} and 1.3×10^{-4} for β values and *M* values, respectively.

Table 3 CpG site-specific regression models before and after normalization step

Values	Methods	Significant sites[b]	CHARGE[c]	Sensitivity	1-Specificity
β values	Standard adjustment[a]	444	357 (80%)	1.9×10^{-2}	2.2×10^{-4}
	Residuals	427	365 (85%)	1.9×10^{-2}	1.5×10^{-4}
	ComBat	600	411 (69%)	2.2×10^{-2}	4.7×10^{-4}
	SVA	96	89 (92%)	0.5×10^{-2}	0.2×10^{-4}
M values	Standard adjustment[a]	322	274 (85%)	1.5×10^{-2}	1.2×10^{-4}
	Residuals	332	299 (90%)	1.6×10^{-2}	0.8×10^{-4}
	ComBat	387	335 (87%)	1.8×10^{-2}	1.3×10^{-4}
	SVA	144	134 (93%)	0.7×10^{-2}	0.2×10^{-4}

Models are adjusted for chip position, recruitment centre, the five percentages of leukocyte subtypes and age at recruitment, menopausal status and BC status
[a]Also adjusted for batch and sample position
[b]Number of significant sites for smoking status after *p* values FDR correction
[c]Number (and percentage) of significant sites identified by the CHARGE meta-analysis

Discussion

Batch effects on DNA methylation measurements have already been documented [13]. Various correcting methods have been recently used, including standard adjustment [3], ComBat [6] and SVA [2]. Our findings suggested that batch was not the only source of variation in the DNA methylation data from our EPIC study, as the position of the sample within the chip and, to a lesser extent, chips within batches, also contributed to total variability. Noteworthy, while variation by batch was essentially random, the position of the sample within the chip contributed systematic variation, with methylation levels progressively increasing by row, but not by column. This might be due to the washing step which is done row by row in each chip during the measurement of DNA methylation using HM450K. Eventually, batch and row sample positions explained cumulatively more than 20% of the methylation levels and were the most important sources of variation. Further replications are needed in others dataset from other labs to validate our findings.

PC-PR2 is a powerful method to identify and quantify random and systematic sources of variation in large-scale datasets. Here, the method, initially developed for metabolomics data [17], was successfully applied to epigenetics data, a challenging set characterized by hundreds of thousands of features, and can easily be extendable to other *-omics* data. It is based on the

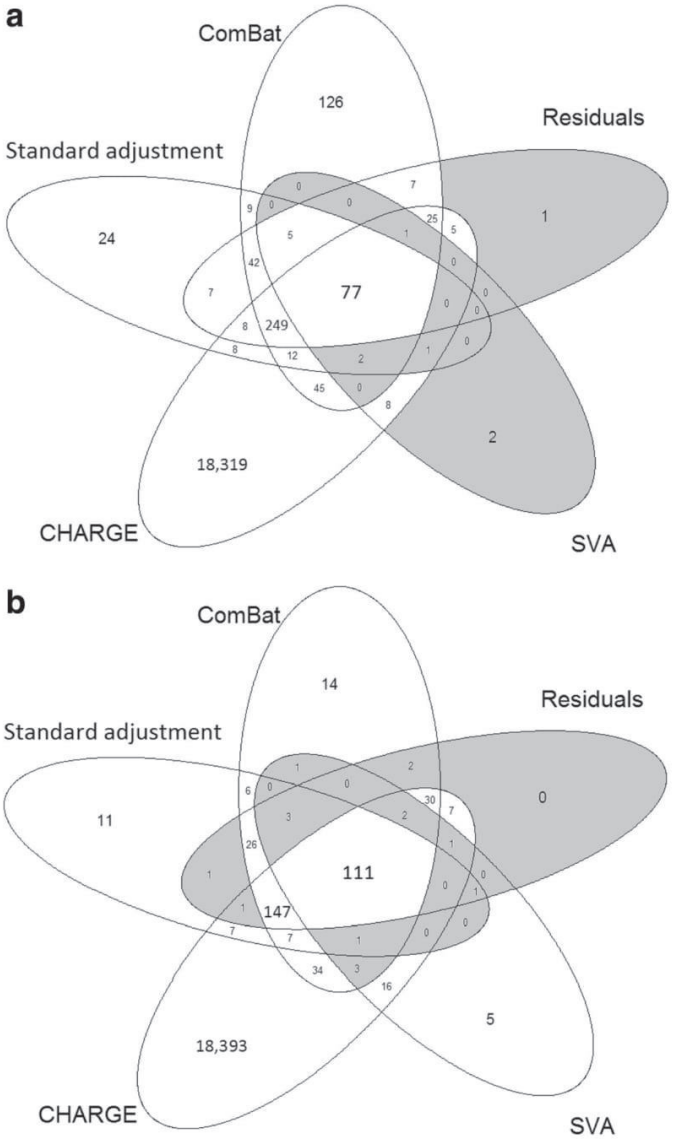


Fig. 4 Venn diagram of significantly identified CpG sites for smoking status using each correcting methods and CHARGE. **a** β values. **b** *M* values. *p* values were corrected for multiple testing with FDR

combination of a principal component analysis (PCA) and the concept of partial R^2 in multivariable linear regression. PC-PR2 quantifies the contribution of variability of continuous and/or categorical covariates to total variability in the outcome data, and in general offers high level of flexibility to capture specific features such as, say, non-linear effects and longitudinal data. A particularly appealing feature is the possibility of performing PCA by decomposing the matrix $XX^{'}$ of dimension $n \times n$ rather than $X^{'}X$ of dimension $p \times p$ that would be virtually untreatable in the *-omics* domain. The PC-PR2 can also be extended to the Infinium MethylationEPIC BeadChip (850K), which is the updated version of HM450K.

Identifying unwanted sources of variation in epigenetics data is a crucial step prior to statistical analysis. Each of the three tested methods succeeded to correct DNA methylation levels for the pre-specified sources of variability.

Percentages of variability due to batch and row sample position diminished to marginal levels after the use of the three methods. Other unknown or unmeasured experimental conditions are also likely to modify DNA methylation measurements, such as differences in sample handling and preparation and the room temperature during sample processing. Overall, the procedures for sample treatment are way more challenging to control, possibly because detailed information on each sample are not always documented, and it is rather assumed that these are relatively homogeneous across recruitment centres. Statistical adjustment for centre is a standard practice in the analysis of epigenetics data and of any laboratory measurements. In this respect, SVA turned out to provide a correction on top of the pre-specified sources of variability through the estimation of surrogate variables possibly influencing overall variability. It was remarkable that the variability attributed to chip position, whose partial R^2 values was 6.5% in the raw data, decreased to 0.9% after SVA, even if chip position was not included in the list of covariates of which we want to remove the variability, specified in the SVA model. Indeed, the surrogate variables, computed by a PCA step in the SVA algorithm, capture the variability in the methylation data which is not already explained by the a priori list of covariates (batch and row sample position). A challenge of DNA methylation data is the presence of outliers that can generate spurious associations. Techniques have been introduced to filter out outliers through preliminary quality control checks globally on all CpG sites [32]. This was achieved through the Illumina GenomeStudio quality in the present study [22]. Nevertheless, outlier values passed the GenomeStudio quality control screening and were detected after applying the residuals or SVA methods. On the contrary, ComBat is based on an empirical Bayesian procedure with an additive and a multiplicative component, the latter contributing to shrink all observations, including outliers [14]. This makes ComBat an attractive solution to control outlier values in large-dimension data. Another interesting feature is that ComBat preserved the observed variability of methylation data in the $[0, 1]$ interval for β values, unlike the residuals' and SVA methods, for which the corrected values could fall outside the $[0, 1]$ range.

The performance of the various correction methods was evaluated in this study through the comparison with results of association between smoking and methylation from the CHARGE consortium, one of the largest studies available to date. This could be a debatable choice but allowed a reference group to be established to compute relative sensitivity and specificity of each normalizing method. The low sensitivity across all methods in our analysis might be explained by the lack of power due to the sample size: over 16,000 samples were included in CHARGE against 902 in our study. Some different characteristics of our population and the one of the CHARGE consortium might also explain the difference in terms of significant sites. For example, only women are included in our analysis and half of them developed latter a breast cancer. This makes more difficult the identification of false positives based on the results from the CHARGE consortium. The analysis showed that ComBat had the highest level of relative sensitivity, i.e. relatively less false negative CpG associated to smoking, compared to the residuals and SVA, consistently for β or M values. On the other hand, SVA came across as the method with, by far, the highest specificity, possibly indicating lesser predisposition to the commit of false positives. As SVA made a much more aggressive correction of systematic variability, the sites identified by SVA are more likely to be universal disruption due to smoking which can explain its higher specificity and its lower sensitivity. In order to avoid over-adjustment using SVA, latent covariates related to subgroups such as the chip position should not be included in the regression model. SVA outperformed both the residuals and, in particular, ComBat, whose lack of specificity turned out to be substantial. In research domains characterized by the danger of populating the scientific literature with false positive findings, like in the *-omics* era, the performance of SVA towards conservative results was deemed to be a valuable feature. Our results would need to be replicated in another dataset.

The β values are approximations of the percentage of methylation in a CpG site. Their distribution is often skewed and ranged from 0 to 1. On the other hand, M values approximate a normal distribution but are more complex to interpret, as they do not have an obvious biological meaning. It has been recommended to use M values for conducting methylation analysis and to use the β values when reporting results due to their intuitive biological interpretation [26]. In our study, the PC-PR2 method identified the same sources of variability explaining a similar amount of the total variability using M or β values. This is likely a consequence on the fact that PC-PR2 is a descriptive method that does not use statistical inference. The association between smoking and DNA methylation was slightly attenuated in terms of number of significant sites using the M values, rather than β values, for the standard adjustment, residuals' correction and ComBat correction. Only SVA identified more significant sites with the M values. β values were more sensitive but less specific than M values, i.e. more significant sites, including both true and false positive sites.

Approaches for correcting batch effects have been compared using microarray data of gene-expression profiles [33]. In that study, a parametric prior ComBat and a non-parametric ComBat were compared to SVA and to three other methods, including distance-weighted discrimination [34], mean-centering [35] and geometric ratio-based [36] methods. Using two microarray datasets

from brain RNA samples and two simulated datasets, ComBat outperformed overall the other methods. In particular, both parametric and non-parametric ComBat algorithms allowed a better control of the variation attributed to batch effect and a better increase of Pearson's correlation coefficient of the replicates in the microarray data and determined the largest AUC in their assessment of overall performance.

ComBat has also been compared to six other methods to correct for batch effect in microarray data [37], including Deming regression [38], Passing-Bablok regression [39], linear mixed model, a third-grade polynomial regression, the non-linear Qspline method [40] and the ReplicateRUV approach [41]. The first five methods calculate residuals based on different regression models. ReplicateRUV removes unwanted variation based on negative control genes and sample replicates. The combination of quantile normalization and ComBat in large-scale gene expression data in the Gutenberg Health Study removed batch effect and preserved biological variability [37].

In this work, we chose to focus on the residuals, ComBat and SVA approaches, because they are the currently most common methods used to remove unwanted variation in DNA methylation. This work can also be applied to the newer methods which are recently available such as the Bacon approach, a Bayesian method to control bias and inflation in EWAS and TWAS based on estimation of the empirical null distribution [42].

Conclusions

Our results suggest that in order to reduce the contribution to systematic variation of DNA methylation, it is essential to randomly allocate samples within chips and batches. This is particularly relevant in nested studies for case-control pairs, possibly within the same row position within a chip. We have shown that the PC-PR2 method on DNA methylation levels lent itself as a very useful tool to explore an a priori list of laboratory factors and sample characteristics and to identify the ones possibly determining unwanted variability in large-scale dimension sets such as epigenetics data. This step turned out to be essential to guide the choice of correcting methods, such as the regression-based residuals, ComBat or SVA, and to further appreciate the extent of these corrections. These steps should be part of the pre-processing analysis of any *-omics* data. SVA should specifically be considered when sources of variability are not known. ComBat and the residuals' method require that potential sources of variability are identified.

Additional files

Additional file 1: Figure S1. Box plots of global methylation (*M* values) according to laboratory factors: batch (a), chip position within batches (b), sample position within chips (c). (PDF 99 kb)

Additional file 2: Figure S2. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the β values. The inflation factor λ is defined as the ratio of the median of the observed $\log_{10}$ transformed p values from the CpG site-specific analysis and the median of the expected $\log_{10}$ transformed p values. (PDF 110 kb)

Additional file 3: Figure S3. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the *M* values. The inflation factor λ is defined as the ratio of the median of the observed $\log_{10}$ transformed p values from the CpG site-specific analysis and the median of the expected $\log_{10}$ transformed p values. (PDF 110 kb)

Abbreviations

BC: Breast cancer; EPIC: European Prospective Investigation into Cancer and nutrition; FDR: False discovery rate; HM450K: Illumina Infinium HumanMethylation450K; PC-PR2: Principal component partial R-square; SVA: Surrogate variables analysis

Acknowledgements

The authors would like to thank the financial support provided by La Fondation de France for a doctoral fellowship. They are also grateful for all the women who participated in the EPIC cohort and without whom this work would not have been possible.

Funding

This work was supported by 'Fondation de France' (2015 00060737) through a doctoral fellow to FP. A grant from the Institut National du Cancer (INCa, France) (2012-070) was awarded to IR and ZH. ZH was also supported by the European Commission (EC) Seventh Framework Programme (FP7) Translational Cancer Research (TRANSCAN) Framework, the Fondation Association pour la Recherche contre le Cancer (ARC, France) and the EC FP7 EurocanPlatform: A European Platform for Translational Cancer Research (grant number: 260791). In addition, this study was supported by postdoctoral fellowship to SA from the International Agency for Research on Cancer, partially supported by the EC FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND). Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten supports SH. AC and KKO are supported by MRC programme grants [MC_UU_12015/1, MC_UU_12015/2 and [MR/L00002/1]. THN is supported by UiT - the Arctic University of Norway. The Hellenic Health Foundation is supporting EPIC-Greece. The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Availability of data and materials

Not applicable.

Authors' contributions

FP performed the statistical data analysis and drafted the manuscript. PF developed the concept of the study with FP, and contributed to draft the manuscript. SA was responsible for the technical aspects of DNA methylation acquisition. IR and ZH conceived the epigenetics study in the nested case-control study on breast cancer, and critically reviewed the manuscript. SA, AK and AN contributed to the interpretation of the results. LB and PV were involved in the data interpretation. All authors contributed to draft the final versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study was approved by the Ethical Review Board of the International Agency for Research on Cancer, and by the local Ethics Committees in the participating centres. This study was also conducted in accordance with the IARC Ethic Committee (Project No 10-22).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Nutritional Methodology and Biostatistics Group, International Agency for Research on Cancer (IARC), World Health Organization, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France. [2]Epigenetics Group, IARC, Lyon, France. [3]Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy. [4]MAP5 – UMR CNRS 8145, Université Paris Descartes, Sorbonne Paris Cité, Paris, France. [5]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden. [7]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK. [8]IIGM – Italian Institute for Genomic Medicine, Torino, Italy. [9]Department of Community Medicine, UiT - The Arctic University of Norway, Tromsø, Norway. [10]Section for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark. [11]Department of Cardiology, Aalborg University Hospital, Aalborg, Denmark. [12]CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine – UVSQ, INSERM, Université Paris-Saclay, Villejuif, France. [13]Gustave Roussy, Villejuif, France. [14]Centre for Health Protection (pb12), National Institute of Public Health and the Environment (RIVM), Bilthoven, Netherlands. [15]Hellenic Health Foundation, Athens, Greece. [16]WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece. [17]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [18]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [19]MRC/PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [20]Nutritional Epidemiology Group, IARC, Lyon, France.

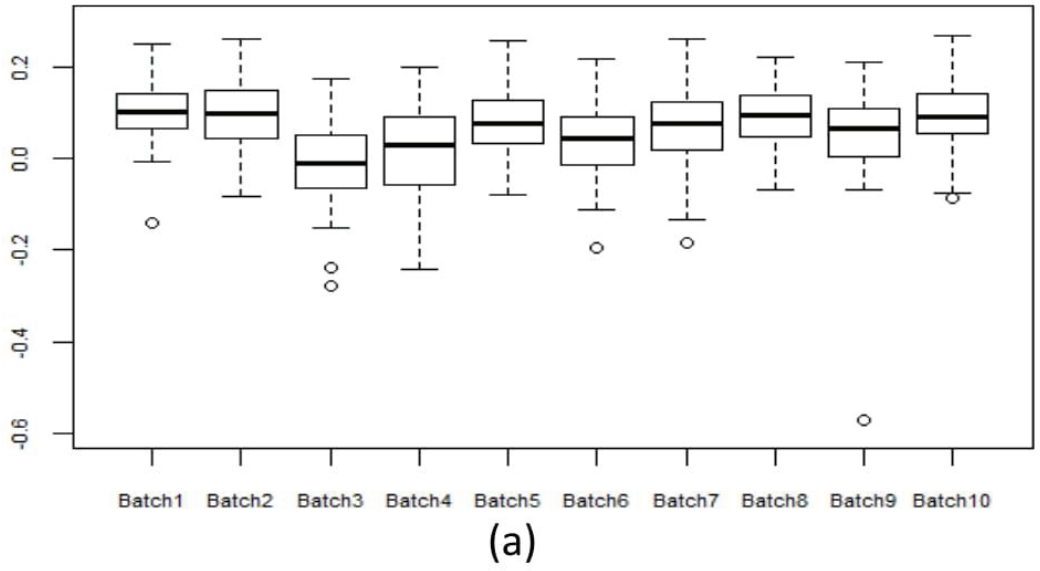
Received: 22 September 2017 Accepted: 12 March 2018
Published online: 21 March 2018

References

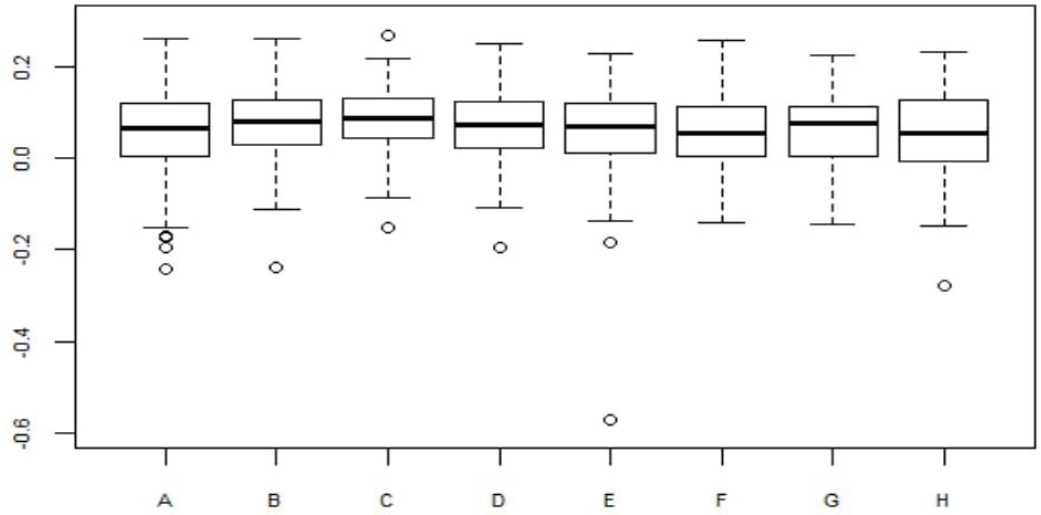
1. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes Dev. 2009;23:781–3.
2. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8:599–618.
3. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet. 2015;24:2349–59.
4. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436–47.
5. Kruman II, Fowler AK. Impaired one carbon metabolism and DNA methylation in alcohol toxicity. J Neurochem. 2014;129:770–80.
6. Joubert BR, den Dekker HT, Felix JF, Bohlin J, Ligthart S, Beckett E, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. Nat Commun. 2016;7:10577.
7. Ba Y, Yu H, Liu F, Geng X, Zhu C, Zhu Q, et al. Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II in maternal and cord blood. Eur J Clin Nutr. 2011;65:480–5.
8. Barrow TM, Michels KB. Epigenetic epidemiology of cancer. Biochem Biophys Res Commun. 2014;455(1-2):70–83. https://doi.org/10.1016/j.bbrc.2014.08.002
9. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95. https://doi.org/10.1016/j.ygeno.2011.07.007
10. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A. 2012;109(26):10522–7. https://doi.org/10.1073/pnas.1120658109
11. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2011;13:97–109.
12. Herceg Z, Ghantous A, Wild CP, Sklias A, Casati L, Duthie SJ, et al. Roadmap for investigating epigenome deregulation and environmental origins of cancer. Int J Cancer. 2018;142(5):874–82. https://doi.org/10.1002/ijc.31014
13. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11 https://doi.org/10.1038/nrg2825.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England). 2007;8:118–27.
15. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3. https://doi.org/10.1371/journal.pgen.0030161
16. Leek JT, Storey JD. A general framework for multiple testing dependence. Proc Natl Acad Sci U S A. 2008;105:18718–23.
17. Fages A, Ferrari P, Monni S, Dossus L, Floegel A, Mode N, et al. Investigating sources of variability in metabolomic data in the EPIC study: the principal component partial R-square (PC-PR2) method. Metabolomics. 2014;10:1074–83.
18. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5:1113–24.
19. Wang SC, Petronis A. DNA methylation microarrays: experimental design and statistical analysis. Boca Raton: Hall; 2008.
20. Matejcic M, de Batlle J, Ricci C, Biessy C, Perrier F, Huybrechts I, et al. Biomarkers of folate and vitamin B12 and breast cancer risk: report from the EPIC cohort. Int J Cancer. 2017;140:1246–59.
21. Ambatipudi S, Horvath S, Perrier F, Cuenin C, Hernandez-Vargas H, Le Calvez-Kelm F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. Eur J Cancer. 2017;75:299–307.
22. Illumina. GenomeStudio/BeadStudio software methylation module. 2011.
23. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.
24. Bolstad BM. Probe level quantile normalization of high density oligonucleotide array data. 2001.
25. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics (Oxford, England). 2013;29:189–96.
26. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
27. 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;13:1–16.
28. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53:325–38.
29. Kleinbaum DG, Kupper LL, Nizam A, Rosenberg ES. Applied regression analysis and other multivariable methods. Nelson Education; 2013.
30. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015;12:115–21.
31. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics (Oxford, England). 2012;28:882–3.
32. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, et al. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109:1394–402.
33. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011;6:e17238.
34. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics (Oxford, England). 2004;20:105–14.
35. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. BMC Med Genet. 2008;1:1–14.
36. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010;10:278–91.
37. Müller C, Schillert A, Röthemeier C, Trégouët D-A, Proust C, Binder H, et al. Removing batch effects from longitudinal gene expression—quantile

normalization plus ComBat as best approach for microarray transcriptome data. PLoS One. 2016;11:e0156594.
38. Martin RF. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clin Chem. 2000;46:100–4.
39. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, part I. Journal of clinical chemistry and clinical biochemistry. Zeitschrift fur klinische Chemie und klinische Biochemie. 1983;21:709–20.
40. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002;3:research0048.
41. Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Biostatistics (Oxford, England). 2016;17:16–28.
42. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017;18:19.

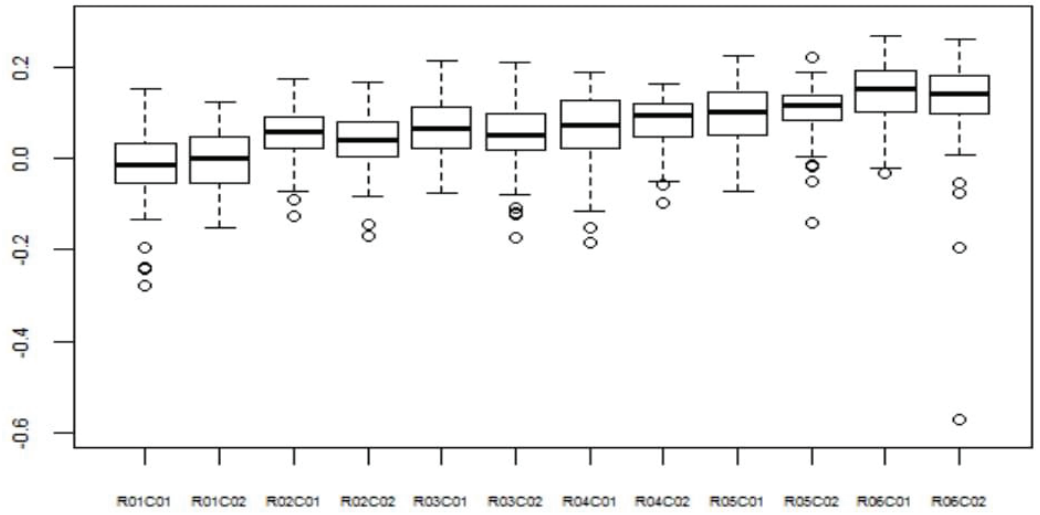
Additional file 1: Figure S1. Box plots of global methylation (M values) according to laboratory factors: batch (a), chip position within batches (b), sample position within chips (c).



(a)

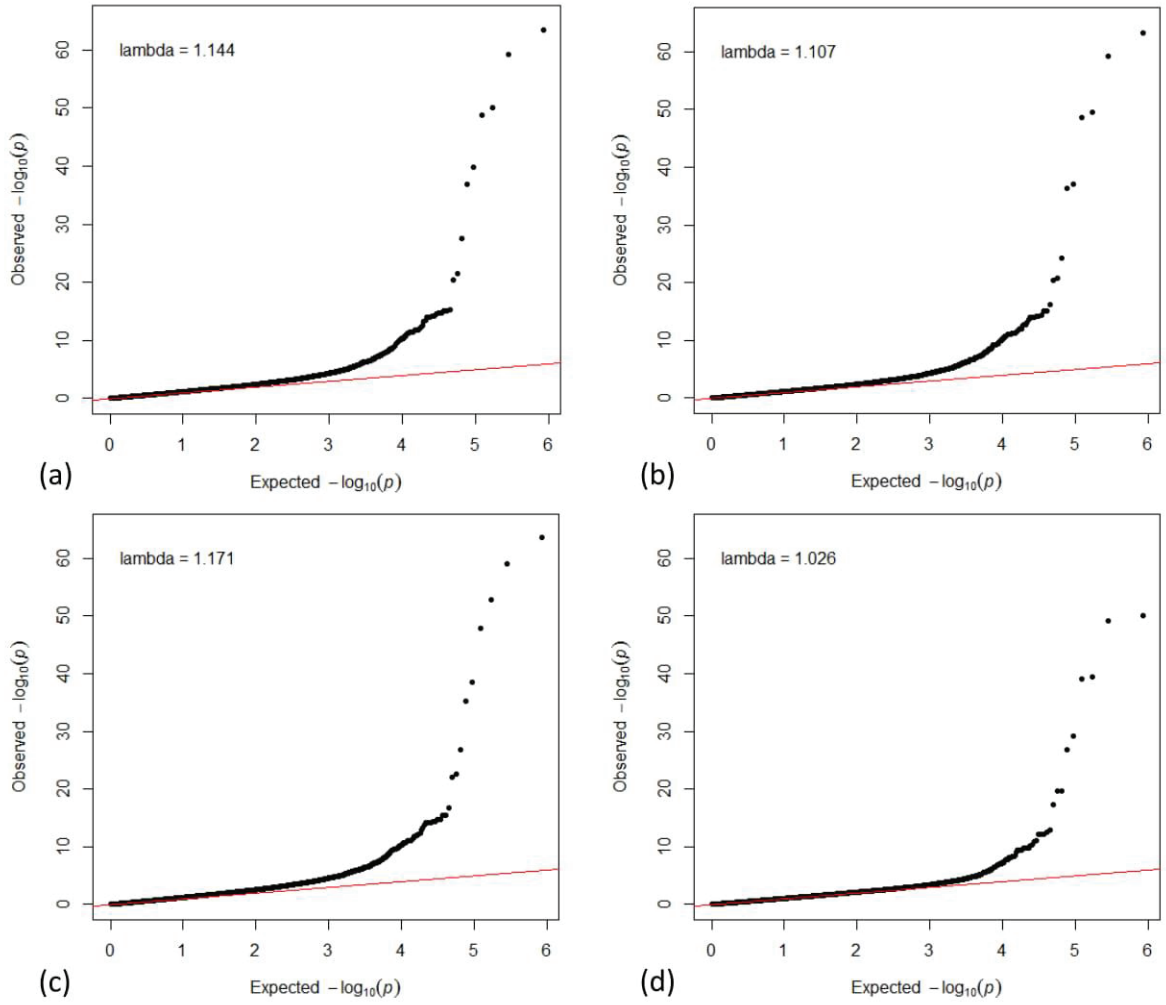


(b)

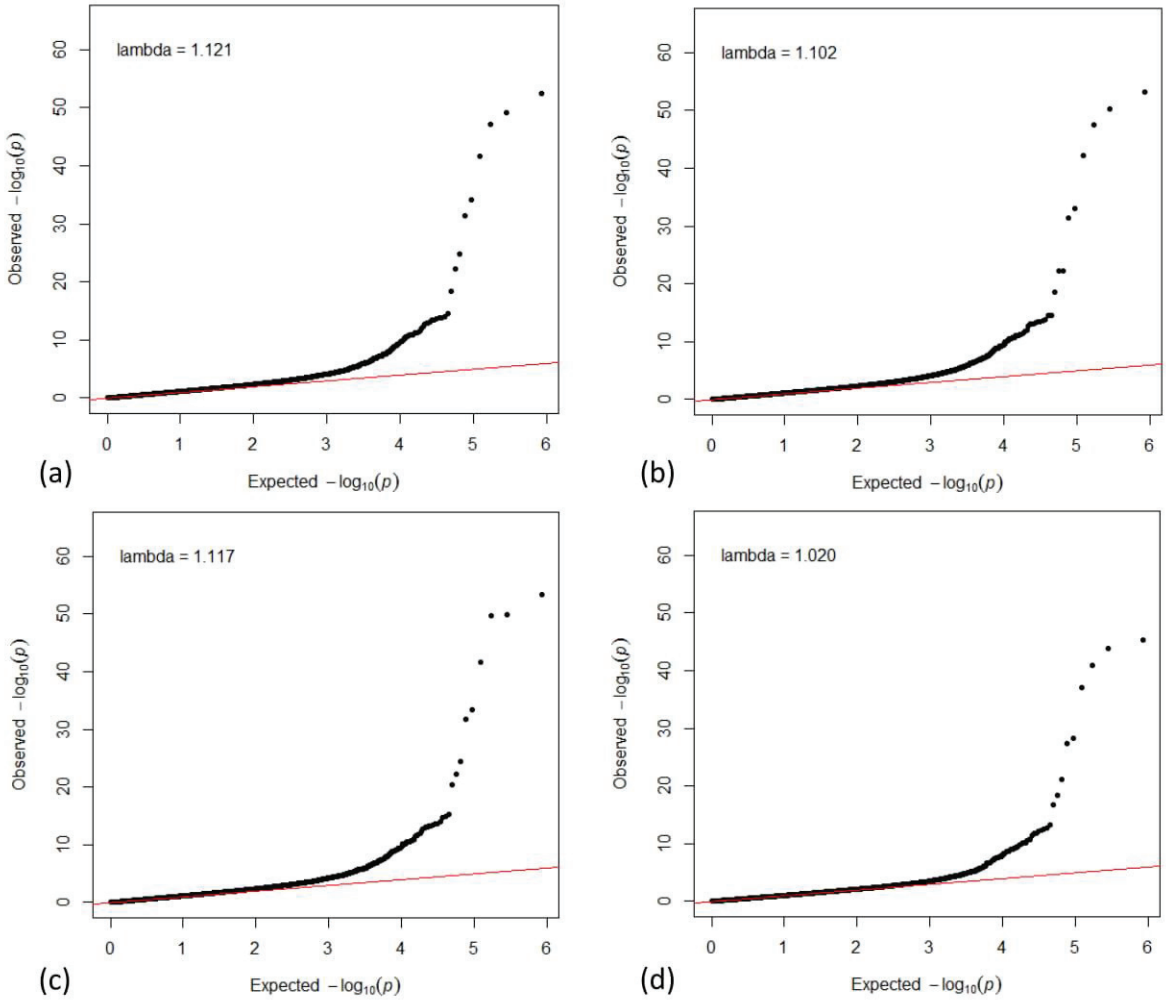


(c)

Additional file 2: Figure S2. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the β values. The inflation factor λ is defined as the ratio of the median of the observed log10 transformed p values from the CpG site-specific analysis and the median of the expected log10 transformed p values.



Additional file 3: Figure S3. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the M values. The inflation factor λ is defined as the ratio of the median of the observed log10 transformed p values from the CpG site-specific analysis and the median of the expected log10 transformed p values.



PART II: Folate, DNA methylation and breast cancer association

1- Association of biomarkers of folate and vitamin B12 with breast cancer risk

Context

Among dietary factors, deficiencies in B vitamins related to Western dietary patterns have been suggested to play a role in breast carcinogenesis (79, 80). Prospective studied, which investigated the effect of biomarkers of vitamin B9 (folate) and vitamin B12 (cobalamin) on BC risk have reported inconsistent findings (70, 81, 82). Blood folate has been inversely associated with BC risk, but a lack of association has also been observed. Similar mixed results have been reported for the association between biomarkers of vitamin B12 and BC risk. A number of factors have been suggested to influence the association between B vitamins and the risk of BC, including menopausal status, alcohol consumption, nutrient interactions and methylenetetrahydrofolate reductase (MTHFR). BC subgroups related to hormone receptor status have been associated with folate intake among premenopausal women.

Objectives

- To evaluate the association between plasma concentrations of folate and vitamin B12 and BC risk overall and stratified by hormone receptor status and potential risk factors in the EPIC cohort.
- To examine the interaction between the MTHFR 677C>T (rs1801133) and 1298A>C (rs1801131) polymorphisms and the two plasma B vitamins on the risk of BC.

Approach

Plasma concentrations of folate and vitamin B12 were determined in 2,491 BC cases individually matched to 2,521 controls among cancer-free women (except non melanoma skin cancer) who provided blood samples at recruitment. Matching criteria included study centre, age at blood donation, exogenous hormone use at blood collection, menopausal status, fasting status and phase of the menstrual cycle at recruitment.

Multivariable logistic regression models were used to estimate odds ratios (OR) by quartiles of either plasma B vitamins. Models were adjusted for BMI, height, alcohol intake, total energy intake, educational attainment, physical activity, ever use of hormone replacement therapy, parity and age at first full-term birth combined and family history of BC. Subgroup

analyses by menopausal status, hormone receptor status of breast tumors (estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2), alcohol intake and MTHFR polymorphisms (677C > T and 1298A > C) were also performed. In addition, the association between each plasma biomarker and the risk of BC was examined using four-knot restricted cubic splines with the midpoint of the fifth decile of plasma vitamin B12 as the reference category. Tests for interaction between each plasma biomarker as continuous variable and potential risk factors were computed by likelihood ratio test.

Main findings

Continuous and quartiles of plasma levels of folate and vitamin B12 were not significantly associated with the overall risk of BC. No further significant association emerged for folate and vitamin B12 after stratification by menopausal status, by hormone receptors status or adjustment for MTHFR polymorphisms.

The interaction term between tertiles of plasma folate (<10.96, 10.96-17.85, >17.85 ηmol/L) and categories of alcohol intake (0-3, 3-12, >12 g/day) was not significantly associated with BC risk ($p_{interaction}$=0.69). Similarly, no significant association between plasma folate and BC risk was observed by median level of alcohol consumption.

A borderline positive association was found between quartiles of vitamin B12 and BC risk in women consuming above the median level of alcohol, i.e. higher than 3.36 g/day, ($OR_{\mathrm{Q4-Q1}}$=1.26; $CI_{95\%}$= [1.00–1.58]; p_{trend}=0.05). BC risk was also significantly increasing according to quartiles of vitamin B12 in women with plasma folate levels below the median value, i.e. lower than 13.56 ηmol/L, ($OR_{\mathrm{Q4-Q1}}$=1.29; $CI_{95\%}$=[1.02–1.62]; p_{trend}=0.03).

Conclusion

Overall, no clear support for an association between plasma levels of folate and BC risk was found in this large prospective study. However, potential interactions between vitamin B12 and alcohol or folate on the risk of BC were observed. Our findings suggest a potential role of vitamin B12 in breast carcinogenesis and raise the possibility of important nutrient–nutrient and gene–nutrient interactions, such as changes in DNA methylation, in the etiology of BC. The potential deleterious effect of high vitamin B12 status in combination with other risk factors for BC deserves further investigation. Given the inconsistent findings to date and the possibility that associations between folate and BC could be influenced by some factors yet to be identified, further studies based on novel biomarkers that take into account the effect of potential risk factors and genetic polymorphisms are warranted.

Published article: Biomarkers of folate and vitamin B12 and breast cancer risk: report from the EPIC cohort.

IJC
International Journal of Cancer

Biomarkers of folate and vitamin B12 and breast cancer risk: report from the EPIC cohort

M. Matejcic[1], J. de Batlle[1†], C. Ricci[1], C. Biessy[1], F. Perrier[1], I. Huybrechts[1], E. Weiderpass[2,3,4,6], M.C. Boutron-Ruault[5,7], C. Cadeau[5,7], M. His[5,7], D.G. Cox[8], H. Boeing[9], R.T. Fortner[10], R. Kaaks[10], P. Lagiou[11,12,13], A. Trichopoulou[11,12], V. Benetou[12], R. Tumino[14], S. Panico[15], S. Sieri[16], D. Palli[17], F. Ricceri[18,19], H.B(as) Bueno-de-Mesquita[20,21,22], G. Skeie[3], P. Amiano[23,24], M.J. Sánchez[23,25], M.D. Chirlaque[23,26,27], A. Barricarte[23,28,29], J.R. Quirós[30], G. Buckland[31], C.H. van Gils[32], P.H. Peeters[32,33], T.J. Key[34], E. Riboli[33], B. Gylling[35], A. Zeleniuch-Jacquotte[36], M.J. Gunter[1], I. Romieu[1] and V. Chajès[1]

[1] International Agency for Research on Cancer, Lyon, France
[2] Genetic Epidemiology Group, Folkhälsan Research Center, Helsinki, Finland
[3] Department of Community Medicine, University of Tromsø – The Arctic University of Norway, Tromsø, Norway
[4] Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway
[5] Institut Gustave Roussy, Villejuif, France
[6] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[7] Université Paris-Saclay, Université Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France
[8] Centre Léon Bérard, INSERM U1052, Cancer Research Center of Lyon, Lyon, France
[9] Epidemiology, German Institute of Human Nutrition Potsdam-Rehbruecke (DIfE), Nuthetal, Germany
[10] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[11] Hellenic Health Foundation, Athens, Greece
[12] WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece
[13] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA
[14] Cancer Registry and Histopathology Unit, Civic - M.P. Arezzo Hospital, ASP Ragusa, Ragusa, Italy
[15] Dipartimento di Medicina Clinica e Chirurgia, Università degli Studi di Napoli Federico II, Naples, Italy
[16] Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy
[17] Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy
[18] Unit of Cancer Epidemiology, Department of Medical Sciences, University of Turin, Turin, Italy

Cancer Epidemiology

Key words: plasma biomarkers, folate, vitamin B12, alcohol, hormone receptor status, MTHFR polymorphism, breast cancer

Abbreviations: BC: breast cancer; BMI: body mass index; CIs: confidence intervals; ER: estrogen receptor; ENDB: EPIC nutrient database; EPIC: European Prospective Investigation into Cancer and Nutrition; FCT: food composition tables; FFQs: food-frequency questionnaires; HER2: human epidermal growth factor receptor 2; ICD: Injuries and Causes of Death; ICC: intraclass correlation coefficient; MTHFR: methylenetetrahydrofolate reductase; ORs: odds ratios; PR: progesterone receptor; QC: quality control; SNPs: single nucleotide polymorphisms; SD: standard deviation; THF: tetrahydrofolate; WCRF: World Cancer Research Fund

[†]J.d.B. is a co-first author

Grant sponsor: International Agency for Research on Cancer; **Grant sponsor:** European Commission FP7 Marie Curie Actions-People-Cofunding of regional, national, and international programs (COFUND); **Grant sponsor:** Ecumenical Project for International Cooperation; **Grant sponsor:** Danish Cancer Society (Denmark); **Grant sponsor:** Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, Institut National de la Santé et de la Recherche Médicale (INSERM) (France); **Grant sponsor:** Deutsche Krebshilfe, Deutsches Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); **Grant sponsor:** Hellenic Health Foundation (Greece); **Grant sponsor:** Italian Association for Research on Cancer (AIRC) and National Research Council (Italy); **Grant sponsor:** Dutch Ministry of Public Health, Welfare and Sports (VWS); **Grant sponsor:** Netherlands Cancer Registry (NKR); **Grant sponsor:** LK Research Funds; **Grant sponsor:** Dutch Prevention Funds; **Grant sponsor:** Dutch ZON (Zorg Onderzoek Nederland); **Grant sponsor:** World Cancer Research Fund (WCRF); **Grant sponsor:** Statistics Netherlands (The Netherlands); **Grant sponsor:** AGAUR, Generalitat de Catalunya; **Grant number:** Exp. 2014 SGR 726; **Grant sponsor:** Health Research Funds; **Grant number:** RD12/0036/0018; **Grant sponsor:** Swedish Cancer Society, Swedish Scientific Council and Regional Government of Skåne and Västerbotten (Sweden); **Grant sponsor:** Cancer Research UK; **Grant sponsor:** Medical Research Council; **Grant sponsor:** Stroke Association; **Grant sponsor:** British Heart Foundation; **Grant sponsor:** Department of Health, Food Standards Agency; **Grant sponsor:** Welcome Trust (United Kingdom); **Grant sponsor:** World Cancer Research Funds; **Grant sponsor:** Institut National du Cancer (INCA); **Grant sponsor:** la Fondation de France (FDF); **Grant sponsor:** La Ligue Nationale contre le Cancer (LNCC)

DOI: 10.1002/ijc.30536

History: Received 5 Aug 2016; Accepted 18 Oct 2016; Online 1 Dec 2016

Correspondence to: Marco Matejcic, International Agency for Research on Cancer, 150, Cours Albert-Thomas, 69372 Lyon CEDEX 08, France, Tel.: +33-0-4-72-73-8029, E-mail: matejcicm@fellows.iarc.fr

[19] Unit of Epidemiology, Regional Health Service ASL TO3, Grugliasco, Italy
[20] Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
[21] Department for Determinants of Chronic Diseases (DCD), National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[22] Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, United Kingdom
[23] CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[24] Public Health Division of Gipuzkoa, BioDonostia Research Institute, San Sebastian, Spain
[25] Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibs, GRANADA, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain
[26] Department of Epidemiology, Regional Health Council, IMIB-Arrixaca, Murcia, Spain
[27] Department of Health and Social Sciences, Universidad de Murcia, Murcia, Spain
[28] Navarra Institute for Health Research (IdiSNA), Pamplona, Spain
[29] Navarra Public Health Institute, Pamplona, Spain
[30] Public Health Directorate, Asturias, Spain
[31] Unit of Nutrition and Cancer, Cancer Epidemiology Research Programme, Catalan Institute of Oncology (ICO-IDIBELL), Barcelona, Spain
[32] Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[33] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, United Kingdom
[34] Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[35] Department of Medical Biosciences, Pathology, Umeå University, Umeå, Sweden
[36] Department of Population Health, NYU School of Medicine, New York, NY

Epidemiological studies have reported inconsistent findings for the association between B vitamins and breast cancer (BC) risk. We investigated the relationship between biomarkers of folate and vitamin B12 and the risk of BC in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort. Plasma concentrations of folate and vitamin B12 were determined in 2,491 BC cases individually matched to 2,521 controls among women who provided baseline blood samples. Multivariable logistic regression models were used to estimate odds ratios by quartiles of either plasma B vitamin. Subgroup analyses by menopausal status, hormone receptor status of breast tumors (estrogen receptor [ER], progesterone receptor [PR] and human epidermal growth factor receptor 2 [HER2]), alcohol intake and *MTHFR* polymorphisms (677C>T and 1298A>C) were also performed. Plasma levels of folate and vitamin B12 were not significantly associated with the overall risk of BC or by hormone receptor status. A marginally positive association was found between vitamin B12 status and BC risk in women consuming above the median level of alcohol ($OR_{Q4\text{-}Q1} = 1.26$; 95% CI 1.00–1.58; $P_{trend} = 0.05$). Vitamin B12 status was also positively associated with BC risk in women with plasma folate levels below the median value ($OR_{Q4\text{-}Q1} = 1.29$; 95% CI 1.02–1.62; $P_{trend} = 0.03$). Overall, folate and vitamin B12 status was not clearly associated with BC risk in this prospective cohort study. However, potential interactions between vitamin B12 and alcohol or folate on the risk of BC deserve further investigation.

What's new?

Does B-vitamin intake play a role in breast cancer (BC) risk? Results have been inconsistent. In this analysis of data from a large, prospective European study, the authors found that, overall, folate and vitamin B12 status were not clearly associated with BC risk. However, the risk did seem to increase somewhat for women who had higher vitamin B12 levels and either low plasma folate or increased alcohol consumption. The authors suggest that this may involve nutrient-nutrient or gene-nutrient interactions, such as changes in DNA methylation, which require further investigation.

The etiology of breast cancer (BC) is complex and results from the combination of lifetime reproductive events, genetics, dietary and lifestyle factors.[1] According to the latest breast cancer report from the World Cancer Research Fund (WCRF), there is novel evidence that alcohol intake and factors that lead to a greater adult attained height are positively associated with postmenopausal and probably also premenopausal BC.[2] Among dietary factors, deficiencies of B vitamins related to Western dietary patterns have been suggested to play a role in breast carcinogenesis.[3,4]

Vitamin B9 (folate) and vitamin B12 (cobalamin) are two water soluble B vitamins involved in one-carbon metabolism,[5] which generates substrates for DNA methylation and DNA synthesis.[6] Thus, deficiencies of these micronutrients may trigger both genetic and epigenetic procarcinogenic processes.[7] Prospective studies that investigated the association between biomarkers of folate and BC risk have reported either an inverse association[8,9] or no association[10,11] overall. The prospective investigation of the relationship between biomarkers of vitamin B12 and BC risk has also produced mixed

results.[8–12] While studies found no evidence for an association between blood levels of vitamin B12 and BC risk in the overall population,[9–11] an inverse association was independently reported among either postmenopausal women[10] or premenopausal women.[9] However, a recent meta-analysis of prospective studies revealed no significant association between biomarkers of vitamin B12 and BC risk in the subgroup analysis by menopausal status.[13]

A number of factors have been suggested to influence the association between B vitamins and the risk of BC, including menopausal status,[11,14–16] alcohol consumption,[15,17] nutrient interactions[18] and methylenetetrahydrofolate reductase (*MTHFR*) gene polymorphisms.[12] *MTHFR* is a key enzyme in one-carbon metabolism where it balances the folate pool between synthesis and methylation of DNA.[19] Although a number of *MTHFR* single nucleotide polymorphisms (SNPs) have been reported in the literature,[20–22] only the C677T and A1298C SNPs have been consistently associated with decreased enzyme activity and reduced plasma folate levels compared with the wild-type genotypes.[23,24] A prospective study found a positive association between plasma folate concentration and postmenopausal BC among carriers of the *MTHFR* 677T allele.[12] Breast tumors are also subdivided into subgroups according to the expression of sex hormone receptors (estrogen receptor [ER], progesterone receptor [PR] and human epidermal growth factor receptor 2 [HER2]), which have been differentially associated with both folate intake[17] and folate status[11] among premenopausal women only.

We conducted a large nested case–control study within European Prospective Investigation into Cancer and Nutrition (EPIC) to evaluate the association between plasma concentrations of folate and vitamin B12 and BC risk overall and stratified by hormone receptor status and potential risk factors. In addition, we examined the interaction between the *MTHFR* 677C > T (rs1801133) and 1298A > C (rs1801131) polymorphisms and the two plasma B vitamins on the risk of BC using data from a subsample of this nested case–control population.

Material and Methods

Study design

The EPIC study is an ongoing multicenter European cohort study designed to investigate the role of dietary habits and lifestyle factors on the incidence of cancer of various sites, including BC.[25] The cohort includes over 521,000 participants recruited between 1992 and 2000 from 23 centers in 10 European countries (Denmark, France, Germany, Greece, Italy, the Netherlands, Norway, Spain, Sweden and UK). Of 367,903 women (age 35–70 years) recruited into the EPIC study, the present analysis excluded women with prevalent cancers at recruitment ($n = 19{,}853$) and missing diagnosis or censoring date ($n = 2{,}892$). A total of 10,713 women with malignant primary BC were identified after a median follow-up of 11.5 years. The follow-up rate was very high (98.5%: 91.4% alive and 7.1% dead) and only 1.5% of women were lost to follow-up.

Details of the recruitment procedures and data collection in the EPIC study have been previously described in details.[26] Briefly, sociodemographic, lifestyle and dietary data were collected at baseline from all the cohort members by administration of country-specific questionnaires. Anthropometric measurements and peripheral blood samples of the participants were also collected. Methods of blood collection, processing and storage are described in details elsewhere.[27] All participants signed an informed consent for the use of their blood samples and data. The study was approved by the Ethical Review Board of the IARC and those of all national recruiting centers.

Selection of study subjects

A nested case–control study was designed among women who provided a blood sample and completed the lifestyle and dietary questionnaires at recruitment. A total of 2,491 BC cases with a confirmed first diagnosis of invasive BC were identified between 1992 and 2010. Each case was individually matched to at least one control subject chosen randomly among cohort women with available blood samples and free of cancer (except nonmelanoma skin cancer) at the time of diagnosis of the corresponding case. Control subjects were matched to cases for study center, age at blood donation (±3 months), exogenous hormone use at blood collection (yes; no; unknown), menopausal status (pre; surgical post; natural post), fasting status (<3, 3–6, >6 hr) and phase of the menstrual cycle (early follicular, late follicular, periovulatory, mid luteal, other luteal) at recruitment.

Dietary and lifestyle data collection

Dietary data were obtained at enrollment using validated country-specific dietary history and food-frequency questionnaires (FFQs), designed to collect local dietary habits of the participants over the preceding year.[26] Dietary intakes of folate and vitamin B12 were estimated using the updated EPIC Nutrient Database (ENDB),[28] following standardization from country-specific food composition tables (FCT) according to Bouckaert's recommendations.[29] Details on dietary assessment have been discussed previously.[17]

Participants also completed a baseline lifestyle questionnaire providing information on anthropometric and sociodemographic characteristics, reproductive history, family history of cancer, physical activity, alcohol use, smoking habits, use of oral contraceptives, hormone replacement therapy and vitamin supplements in the year prior to enrollment date.

Outcome assessment

Participants were followed from the date of enrollment until first cancer diagnosis, death, emigration or end of the follow-up period, whichever occurred first. Incident cancer cases were identified through population cancer registries (Denmark, Italy except Naples, the Netherlands, Norway, Spain, Sweden and UK) or by a combination of methods including health insurance, cancer and pathology registries and active

follow-up through study subjects and their next-of-kin in three countries (France, Germany, Greece and Naples). Data on clinical and tumor characteristics were coded according to the 10th Revision of the International Statistical Classification of Diseases, Injuries and Causes of Death (ICD).

In the present study, 91% of BC cases were confirmed by histological or cytological examination, whereas the remaining 9% was diagnosed through clinical observation, ultrasound, autopsy or death certificate. The most frequent subtype of BC was ductal carcinoma (71.5%), followed by lobular carcinoma (14.1%) and tubular carcinoma (2.7%). The remaining BC cases were classified as mixed (5.0%) or other (6.7%) subtypes.

Hormone receptor status determination

Determination of ER, PR and HER2 status of BC cases was performed within each EPIC center. Information on hormone receptor status as well as on the methods for its determination was retrieved from each EPIC center using the same approaches used for collection of incident cases. To standardize the quantification of the receptor status collected across centers, the following criteria were applied for a positive receptor status: $\geq$10% cells stained, any "plus-system" description, $\geq$20 fmol/mg, an Allred score of $\geq$3, an IRS $\geq$2 or an H-score $\geq$10.[30] ER, PR and HER2 status was available for 98, 84 and 44% of cases, respectively. For the remaining cases, hormone receptor status was not determined because of insufficient amount of tumor tissue available for histopathological evaluation. Furthermore, HER2 status could not be ascertained in the majority of cases because of the lack of a specific test in the nineties.

Laboratory measurements

All biochemical analyses were performed at the Bevital AS laboratory in Bergen, Norway (www.bevital.no). Microbiological assays were used to determine plasma concentrations of folate[31] and vitamin B12.[32] The assays were adapted to a microtiter plate format and carried out by a robotic workstation. Throughout all steps of the biochemical analysis, samples from each case–control set were analyzed within the same batch. The laboratory personnel were blinded to case–control status. To assess the measurement precision, each batch contained six quality control (QC) samples with known biomarker concentrations and four samples without biomarker (blanks). The six QC samples were three samples in parallels. The coefficient of variation calculated from the three duplicate sets of identical QC samples was 8.6% for folate and 5.0% for vitamin B12. Plasma concentrations of folate and vitamin B12 were determined for all study participants.

Genotyping analysis

Determination of the genotype status was carried out only in a subsample of 401 cases and 401 matched control individuals from this nested case–control population. DNA extraction from white blood cells was carried out using Autopure LS kit (Gentra Systems, Minneapolis, MN). DNA concentration was quantified with Quant-iT PicoGreen dsDNA reagent (Thermo Fisher Scientific, Waltham, MA).

The *MTHFR* 677C > T (rs1801133) and 1298A > C (rs1801131) single nucleotide polymorphisms (SNPs) were genotyped by Kaspar allelic discrimination assay using allele-specific probes and fluorescent reporters (LGC Group, UK). Each reaction was carried out according to the manufacturer's instructions using supplied kits. Amplifications and end-point allele determination were performed in 96-well plates using a StepOne Plus system (Applied Biosystems). Each plate contained randomly placed case and control samples, while matched sets were analyzed within the same plate. Genotyping success rates were 98.0 and 96.5% for rs1801131 and rs1801133, respectively. Samples not yielding genotypes were removed from further analyses.

Statistical methods

Lifestyle and dietary baseline characteristics of study participants were described using mean $\pm$ standard deviation (SD) for continuous variables and percentages for categorical variables. Plasma concentrations of folate and vitamin B12 were log natural transformed to normalize their distribution. The paired t test and χ^2 test were used to assess differences between cases and control individuals with regard to continuous and categorical variables, respectively.

Multivariable conditional logistic regression models were used to estimate odds ratios (ORs) and 95% confidence intervals (95% CIs) for overall BC and specific subgroups stratified by menopausal status at recruitment (dichotomized as natural/surgical postmenopausal and premenopausal) and by hormone receptor status (ER+/ER−, PR+/PR−, HER2+/HER2−). Crude ORs were also presented to observe the effect of confounding on the risk estimates. In addition, the association between each plasma biomarker and the risk of BC was examined using four-knot restricted cubic splines with the midpoint of the fifth decile of plasma vitamin B12 as the reference category.[33]

Quartiles and tertiles of plasma levels of biomarkers for the overall and hormone receptor-specific analyses, respectively, were determined on the basis of the distribution among control individuals. Tests for linear trends were performed by entering the median value of each category as continuous term in the multivariable models.

All multivariate models were adjusted by BMI, height, alcohol intake, total energy intake, educational attainment (primary school, technical/professional school, secondary school, university degree, 4.2% unknown), physical activity (inactive, moderately inactive, moderately active, active, 6.9% unknown), ever use of hormone replacement therapy (never, ever, 4.1% unknown), parity and age at first full-term birth combined (nulliparous, <30 year and 1–2 children, <30 year and $\geq$3 children, >21–30 year, $\geq$30 year, 3.6% unknown) and family history of BC (yes, no, 53.1% unknown). These confounders were previously related to BC risk or blood

measurements and were chosen based on previous studies in the literature. Unknown categories of the above mentioned variables were included in the model using indicator variables.

Multivariate unconditional logistic regression models were used to investigate the association between plasma concentrations of folate and vitamin B12 and BC risk by levels of alcohol intake or plasma folate (low and high levels based on median values) and by *MTHFR* genotypes. The joint effect of plasma folate (in tertiles) and categories of alcohol intake (0–3, 3–12, >12 g/day) on BC risk was evaluated by using the lowest tertile of plasma folate and highest category of alcohol intake as reference category, as previously assessed.[17]

Tests for interaction between each plasma biomarker as continuous variable and potential risk factors were computed by likelihood ratio test. Formal tests of heterogeneity between ORs in menopausal and hormone receptor subgroups were based on χ^2 statistics, calculated as the deviations of logistic beta-coefficients observed in each of the subgroups relative to the overall beta-coefficient.

The association between the SNPs and overall BC risk was evaluated by conditional logistic regression. Genotypic (codominant) and dominant models were assumed for SNP effects. A trend test was conducted by treating the genotypes as equally spaced integer weights and entering the variable as a continuous term in the model.

Specific sensitivity analyses were carried out by excluding women consuming multivitamin supplements and cases diagnosed within the first 2 years of follow-up (to reduce the chance of reverse causality).

Statistical tests were two-sided, and p values below 0.05 were considered statistically significant. All analyses were performed using STATA 12.1 (StataCorp. 2011, Stata Statistical Software: Release 11, College Station, TX).

Results

Table 1 summarizes the sociodemographic, reproductive and lifestyle characteristics of study participants by case–control status. Cases had slightly older age at menopause ($p = 0.026$) and at first live birth ($p = 0.023$) than control individuals. A slightly higher BMI in cases compared with the control group was found among postmenopausal women ($p < 0.001$), but not among premenopausal women. Cases were also more likely to have had a first-degree relative with BC ($p = 0.009$), and had higher daily alcohol intake ($p = 0.002$). Both *MTHFR* SNPs were in Hardy-Weinberg equilibrium ($p = 0.298$ for C677T; $p = 0.823$ for A1298C) and the frequency of the minor allele among control individuals was 30.8% at locus C677T and 37.0% at locus A1298C (data not shown).

There was no significant association between plasma levels of folate and vitamin B12 and the overall risk of BC (Table 2) or by ER, PR and HER2 status (Table 3). No further association emerged after adjustment by *MTHFR* polymorphisms for the available subsample (data not shown).

A nonlinear modeling of the association between plasma concentrations of vitamin B12 and BC risk showed a borderline significant trend ($P_{trend} = 0.07$) in increased risk associated with plasma concentrations of vitamin B12 higher than 360 pmol/l, while the odds ratio plateaued at levels ≥500 pmol/l. No dose-dependent effect of plasma folate on the risk of BC was observed (data not shown).

Because of the impaired folate absorption and altered one-carbon metabolism due to chronic alcohol consumption,[34] we reported risk estimates by tertiles of plasma folate and categories of alcohol consumption (Fig. 1). The association between plasma folate concentration and BC risk was not significantly modified by levels of alcohol intake ($P_{interaction} = 0.69$). Similarly, no significant association between plasma folate and BC risk was observed by median level of alcohol consumption (data not shown).

The association between plasma levels of vitamin B12 and BC risk stratified by the median intake of alcohol is summarized in Table 4. There was a borderline significant increase in risk associated with the highest quartile of plasma vitamin B12 in women consuming at least 3.36 g/day of alcohol ($OR_{Q4\text{-}Q1} = 1.26$; 95% CI 1.00–1.58; $P_{trend} = 0.05$), while no significant association emerged in women drinking lower amounts of alcohol ($OR_{Q4\text{-}Q1} = 1.08$; 95% CI 0.86–1.35; $P_{trend} = 0.56$). However, no significant heterogeneity by alcohol intake was found ($P_{heterogeneity} = 0.14$). The multivariable risk estimates did not change appreciably after further adjustment by plasma folate concentration (data not shown).

A statistically significant interaction between plasma concentrations of folate and vitamin B12 on the risk of BC was observed ($P_{interaction} = 0.04$; data not shown). To further explore this interaction, a stratification analysis by the median level of plasma folate was carried out (Table 4). A marginally increased risk of BC associated with increasing concentrations of plasma vitamin B12 was found in women with plasma levels of folate below 13.56 nmol/l ($OR_{Q4\text{-}Q1} = 1.29$; 95% CI 1.02–1.62; $P_{trend} = 0.03$), while no significant association occurred in women with higher levels of plasma folate ($P_{trend} = 0.68$). A borderline significant heterogeneity by plasma folate levels was also found ($P_{heterogeneity} = 0.05$).

Exclusion from analyses of women who consumed multivitamin supplements or cases diagnosed within the first two years of follow-up did not change the risk estimates in our study population (data not shown).

The *MTHFR* 677C > T and 1298A > C SNPs were in low linkage disequilibrium among both the cases ($r^2 = 0.24$) and control individuals ($r^2 = 0.25$). There was no significant association between either C677T ($OR_{TTvs.CC} = 0.71$; 95% CI 0.42–1.19; $P_{trend} = 0.38$) or A1298C ($OR_{CCvs.AA} = 0.97$; 95% CI 0.62–1.53; $P_{trend} = 0.91$) and the overall risk of BC. The interaction between plasma folate or vitamin B12 and *MTHFR* SNPs was not statistically significant ($P_{interaction} > 0.05$). Plasma concentrations of the two B vitamins were not significantly associated with BC risk in any of the genotypic classes (homozygous wild-type, heterozygous, homozygous variant) of each

Table 1. Characteristics of study population[1]

	Cases (*n*)	Controls (*n*)	*p* difference[2]
No. of individuals, *n* (%)	2491 (49.7%)	2521 (50.3%)	
Mean age (year) at			
Blood collection	54.1 ± 8.4	54.1 ± 8.4	
Diagnosis	60.2 ± 8.8		
Menopause	49.1 ± 4.7	48.7 ± 5.0	0.026
Menarche	13.0 ± 1.5	13.1 ± 1.6	0.010
Age at first birth and parity, *n* (%)			0.023
Nulliparous	349 (14.5)	318 (13.1)	
First birth before age 30 years, 1-2 children	1,086 (45.2)	1,129 (46.5)	
First birth before age 30 years, ≥3 children	590 (24.6)	652 (26.8)	
First birth after age 30 years	376 (15.7)	329 (13.5)	
Unknown≠	90 (3.6)	93 (3.7)	
Menopausal status, *n* (%)			
Premenopause	761 (30.6)	770 (30.5)	
Postmenopause	1,642 (65.9)	1,665 (66.1)	
Perimenopause	88 (3.5)	86 (3.4)	
Ever use of menopausal hormones, *n* (%)			0.820
No	1,687 (70.6)	1,700 (70.4)	
Yes	703 (29.4)	714 (29.6)	
Unknown≠	101 (4.0)	107 (4.2)	
Ever use of contraceptive pill, *n* (%)			0.567
No	1,136 (46.2)	1,166 (46.8)	
Yes	1,325 (53.8)	1,325 (53.2)	
Unknown≠	30 (1.2)	30 (1.2)	
Anthropometric measures			
Adult weight (kg)	66.5 ± 11.7	65.2 ± 11.1	<0.001
Adult height (cm)	161.7 ± 6.5	161.3 ± 6.5	0.009
BMI in premenopause	24.6 ± 4.0	24.6 ± 4.1	0.699
BMI in postmenopause	26.0 ± 4.5	25.4 ± 4.1	<0.001
Waist/Hip Ratio (WHR)	0.792 ± 0.068	0.791 ± 0.066	0.552
Physical activity, *n* (%)			0.260
Inactive	333 (14.3)	293 (12.5)	
Moderately inactive	736 (31.7)	742 (31.6)	
Moderately active	1,073 (46.2)	1,109 (47.3)	
Active	180 (7.7)	201 (8.6)	
Unknown≠	169 (6.8)	176 (7.0)	
Alcohol intake, *n* (%)			0.002
Nondrinkers	440 (17.7)	458 (18.2)	
>0–3 g/day	716 (28.7)	777 (30.8)	
>3–12 g/day	658 (26.4)	713 (28.3)	
>12 g/day	573 (22.7)	677 (27.2)	
Family history of breast cancer, *n* (%)			0.009
No	998 (86.3)	1,071 (89.8)	
Yes	159 (13.7)	122 (10.2)	

Table 1. Characteristics of study population (Continued)

	Cases (*n*)	Controls (*n*)	*p* difference[2]
Unknown≠	1,334 (53.5)	1,328 (52.7)	
Smoking status, *n* (%)			0.752
Never	1,432 (58.8)	1,473 (59.5)	
Former	580 (23.8)	571 (23.0)	
Current	423 (17.4)	433 (17.5)	
Unknown≠	56 (2.2)	44 (1.8)	
Level of education, *n* (%)			0.090
Low	852 (35.6)	883 (36.6)	
Medium	998 (41.7)	1,049 (43.5)	
High	541 (22.6)	479 (19.9)	
Unknown≠	100 (4.0)	110 (4.4)	
Dietary intake			
Energy intake (kcal)	1972.4 ± 549.8	1953.4 ± 555.0	0.210
Dietary folate (μg)	295.6 ± 112.1	296.5118.2	0.681
Dietary vitamin B12 (μg)	6.1 ± 3.5	6.2 ± 3.7	0.306
Vitamin supplement use, *n* (%)			0.870
No	878 (77.0)	879 (77.3)	
Yes	262 (23.0)	258 (22.7)	
Unknown≠	1,351 (54.2)	1,384 (54.9)	
Plasma concentrations			
Folate (nmol/L)[3]	14.1 ± 1.7	14.3 ± 1.8	0.512
Vitamin B12 (pmol/L)[3]	374.2 ± 1.5	370.0 ± 1.5	0.242
***MTHFR* C677T**			0.337
C/C	197 (49.1)	194 (48.4)	
C/T	163 (40.7)	160 (39.9)	
T/T	29 (7.2)	42 (10.5)	
Unknown≠	12 (3.0)	5 (1.2)	
***MTHFR* A1298C**			0.835
A/A	147 (36.7)	154 (38.4)	
A/C	188 (46.8)	178 (44.4)	
C/C	52 (13.0)	54 (13.5)	
Unknown≠	14 (3.5)	15 (3.7)	

[1]Data are presented as means (±SD) or percentages. Geometric means (±SD) of plasma folate and vitamin B12 are presented. Missing values are excluded from calculations.
[2]Statistical significance for differences between cases and controls was tested using paired *t* test for continuous variables and paired χ^2 test for categorical variables.
[3]Differences in plasma concentration of folate and vitamin B12 were assessed on log natural transformed data. For all other variables, differences were assessed on crude data.

SNP (data not shown). No further association emerged in the dominant models, and adjustment for the alternative SNP did not change the risk estimates.

Discussion

In this large prospective European study, circulating levels of folate and vitamin B12 were not significantly associated with the overall risk of BC. However, we found borderline positive associations between plasma concentrations of vitamin B12 and BC risk restricted to women with either high alcohol intake or low folate status. The *MTHFR* C677T and A1298C polymorphisms had no effect modification on the association between either plasma B vitamin and BC risk in a subsample of this nested case–control study.

A study was recently conducted to assess the reliability of plasma biomarkers involved in one-carbon metabolism in a subsample from the EPIC study (38 men and 35 women), which was estimated over a period of 2–5 years using an

Table 2. Crude and multivariable odds ratios[1] for association of plasma folate and vitamin B12 with breast cancer risk overall and stratified by menopausal status at recruitment[2]

Plasma concentration	Matched cases/ controls (*n*)[3]	Crude OR	Multivariable OR[1] (95% CI)	P_{trend}[4]	$P_{heterogeneity}$[5]
Folate (nmol/l)					
All women					
Continuous[6]	2,491/2,521	0.96	0.93 (0.83, 1.05)		
<9.82	624/631	1 (ref)	1 (ref)	0.80	
9.82–13.56	595/630	0.95	0.97 (0.82, 1.15)		
13.56–19.80	663/631	1.06	1.07 (0.90, 1.28)		
>19.80	609/629	0.98	0.94 (0.79, 1.13)		
Menopausal status at recruitment[6]					
Premenopausal women					0.67
Continuous[6]	736/747	0.98	0.99 (0.79, 1.23)		
<9.82	218/220	1 (ref)	1 (ref)	0.61	
9.82–13.56	168/192	0.88	0.88 (0.65, 1.20)		
13.56–19.80	201/179	1.16	1.27 (0.93, 1.75)		
>19.80	149/156	0.97	1.00 (0.72, 1.41)		
Postmenopausal women					
Continuous[6]	1,615/1,634	0.97	0.93 (0.81, 1.07)		
<9.82	385/391	1 (ref)	1 (ref)	0.46	
9.82–13.56	393/406	0.98	1.01 (0.81, 1.26)		
13.56–19.80	418/408	1.04	1.01 (0.81, 1.25)		
>19.80	419/429	0.99	0.94 (0.75, 1.17)		
Vitamin B12 (pmol/l)					
All women					
Continuous[6]	2,489/2,519	1.09	1.10 (0.94, 1.29)		
<293.6	613/630	1 (ref)	1 (ref)	0.24	
293.6–373.1	628/630	1.03	1.00 (0.85, 1.19)		
373.1–460.0	578/630	0.95	0.95 (0.80, 1.13)		
>460.0	670/629	1.13	1.14 (0.95, 1.36)		
Menopausal status at recruitment[6]					0.68
Premenopausal women					
Continuous[6]	735/746	1.01	1.06 (0.78, 1.45)	0.10	
<293.6	181/195	1 (ref)	1 (ref)		
293.6–373.1	176/191	1	0.98 (0.71, 1.35)		
373.1–460.0	187/171	1.22	1.23 (0.90, 1.71)		
>460.0	191/189	1.15	1.26 (0.90, 1.77)		
Postmenopausal women					
Continuous[6]	1,614/1,633	1.13	1.15 (0.95, 1.39)	0.46	
<293.6	407/413	1 (ref)	1 (ref)		
293.6–373.1	421/408	1.05	1.00 (0.81, 1.23)		
373.1–460.0	356/412	0.88	0.88 (0.71, 1.09)		
>460.0	430/400	1.12	1.11 (0.89, 1.39)		

[1]Subjects were matched by study center, age, menopausal status, exogenous hormone use, fasting status and phase of the menstrual cycle. Models were adjusted by date at blood collection, education, BMI, height, physical activity, ever use of hormone replacement therapy, alcohol intake, parity and age at first full-term birth combined, total energy intake and family history of breast cancer.
[2]Menopausal status at recruitment dichotomized as natural/surgical postmenopausal and premenopausal.
[3]Cut points of quartiles determined on control individuals.
[4]Obtained by modeling the median value of tertiles as continuous term in the multivariable model.
[5]Tests of heterogeneity between ORs in menopausal subgroups based on χ^2 statistics calculated as the deviations of logistic beta-coefficients observed in each of the subgroups (premenopausal and postmenopausal women) relative to the overall beta-coefficient.
[6]The OR (95% CI) in the continuous model corresponds to an increment of 2.7 units of folate (nmol/l) or vitamin B12 (pmol/l).

Table 3. Crude and multivariable odds ratios[1] for association of plasma folate and vitamin B12 with breast cancer risk according to hormone receptor status[2]

Plasma concentration	Matched cases/ controls (n)[3]	Crude OR	Multivariable OR[1] (95% CI)	P_{trend}[4]	$P_{heterogeneity}$[5]
Folate (nmol/l)					
ER+					0.63
Continuous[6]	1,987/2,009	0.99	0.96 (0.85, 1.09)		
<10.96	630/677	1 (ref)	1 (ref)	0.51	
10.96–17.85	674/662	1.11	1.11 (0.93, 1.31)		
>17.85	683/670	1.12	1.07 (0.90, 1.27)		
ER−					
Continuous[6]	455/463	0.86	0.89 (0.67, 1.18)		
<10.96	162/153	1 (ref)	1 (ref)	0.55	
10.96–17.85	160/159	0.94	0.97 (0.68, 1.39)		
>17.85	133/151	0.82	0.89 (0.62, 1.29)		
PR+					0.93
Continuous[6]	1,407/1,452	1.01	0.98 (0.84, 1.15)		
<10.96	482/509	1 (ref)	1 (ref)	0.67	
10.96–17.85	482/486	1.06	1.05 (0.85, 1.28)		
>17.85	443/430	1.11	1.05 (0.85, 1.30)		
PR−					
Continuous[6]	690/696	0.96	0.97 (0.77, 1.22)		
<10.96	219/236	1 (ref)	1 (ref)	0.91	
10.96–17.85	245/220	1.21	1.22 (0.91, 1.64)		
>17.85	226/240	1.02	1.03 (0.76, 1.38)		
HER2+					0.71
Continuous[6]	250/252	1.01	1.07 (0.72, 1.60)		
<10.96	66/80	1 (ref)	1 (ref)	0.63	
10.96–17.85	98/78	1.54	1.38 (0.81, 2.35)		
>17.85	86/94	1.13	1.16 (0.68, 1.99)		
HER2−					
Continuous[6]	854/862	1.04	0.98 (0.80, 1.20)		
<10.96	294/314	1 (ref)	1 (ref)	0.43	
10.96–17.85	287/298	1.05	1.10 (0.85, 1.43)		
>17.85	273/250	1.2	1.12 (0.85, 1.47)		
Vitamin B12 (pmol/l)					
ER+					0.40
Continuous[6]	1,986/2,008	1.05	1.06 (0.89, 1.26)		
<323.1	693/684	1 (ref)	1 (ref)	0.54	
323.1–426.0	607/668	0.9	0.90 (0.76, 1.06)		
>426.0	686/657	1.05	1.06 (0.89, 1.26)		
ER−					
Continuous[6]	454/462	1.22	1.26 (0.86, 1.86)		
<323.1	140/147	1 (ref)	1 (ref)	0.26	
323.1–426.0	150/154	1.05	1.16 (0.79, 1.68)		
>426.0	164/162	1.09	1.26 (0.85, 1.86)		

Table 3. Crude and multivariable odds ratios for association of plasma folate and vitamin B12 with breast cancer risk according to hormone receptor status (Continued)

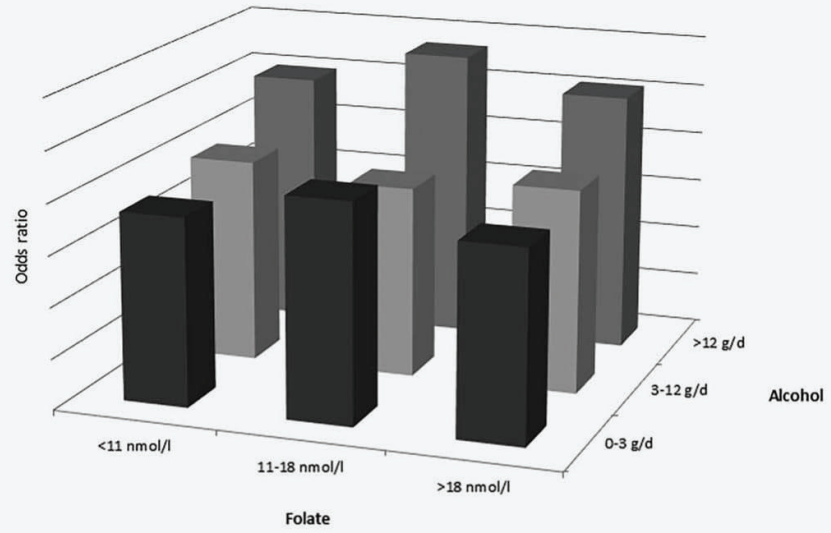
Plasma concentration	Matched cases/ controls (*n*)[3]	Crude OR	Multivariable OR[1] (95% CI)	P_{trend}[4]	$P_{heterogeneity}$[5]
PR+					0.43
Continuous[6]	1,406/1,424	1.02	1.02 (0.82, 1.27)		
<323.1	493/475	1 (ref)	1 (ref)	0.89	
323.1–426.0	434/485	0.86	0.88 (0.72, 1.07)		
>426.0	479/465	1.01	1.02 (0.83, 1.25)		
PR−					
Continuous[6]	689/695	1.13	1.18 (0.88, 1.60)		
<323.1	212/223	1 (ref)	1 (ref)	0.30	
323.1–426.0	225/231	1.05	0.97 (0.72, 1.32)		
>426.0	252/242	1.12	1.18 (0.86, 1.62)		
HER2+					0.98
Continuous[6]	249/251	1.15	1.13 (0.70, 1.83)		
<323.1	85/82	1 (ref)	1 (ref)	0.90	
323.1–426.0	84/83	0.97	1.02 (0.60, 1.74)		
>426.0	80/87	0.87	1.03 (0.59, 1.81)		
HER2−					
Continuous[6]	853/861	1.04	1.12 (0.84, 1.49)		
<323.1	328/311	1 (ref)	1 (ref)	0.66	
323.1–426.0	244/281	0.82	0.85 (0.66, 1.10)		
>426.0	281/270	1	1.08 (0.82, 1.42)		

[1]Subjects were matched by study center, age, menopausal status, exogenous hormone use, fasting status and phase of the menstrual cycle. Models were adjusted by date at blood collection, education, BMI, height, physical activity, ever use of hormone replacement therapy, alcohol intake, parity and age at first full-term birth combined, total energy intake and family history of breast cancer.
[2]Classes of hormone receptors investigated: estrogen receptor positive/negative (ER±), progesterone receptor positive/negative (PR±) and human epidermal growth factor receptor 2 positive/negative (HER2±).
[3]Cut points of tertiles determined on all control individuals.
[4]Obtained by modeling the median value of tertiles as continuous term in the multivariable model.
[5]Tests of heterogeneity between ORs in hormone receptor subgroups based on χ^2 statistics calculated as the deviations of logistic beta-coefficients observed in each of the subgroups (*i.e.*, ER+ and ER− status) relative to the overall beta-coefficient.
[6]The OR (95% CI) in the continuous model corresponds to an increment of 2.7 units of folate (nmol/l) or vitamin B12 (pmol/l).

intraclass correlation coefficient (ICC).[35] The study showed that plasma vitamin B12 was a highly reliable biomarker (ICC = 0.75), while a modest reliability was observed for plasma folate (ICC = 0.45). Because our study was performed on a larger number of subjects and extended over 18 years of follow-up, it is difficult to predict whether and to what extent the single biomarker measurements may have led to attenuation of the risk estimates in our study. However, when the models were adjusted for the regression dilution using the ICCs as adjustment coefficients, no significant change in risk estimates was observed.

Consistent with our findings, a prospective study within EPIC reported a lack of significant association between dietary folate intake and the overall risk of BC.[17] Prospective investigations based on biomarkers of nutrient status reported inconsistent findings between folate and BC risk.[8–12] The mean plasma folate concentration in our study population (14.2 nmol/l) was comparable to that reported in the Malmö Diet and Cancer cohort (12.8 nmol/l), which also reported a null association between plasma levels of folate and overall BC risk.[12] However, our highest category of plasma folate (>19.8 nmol/l; 609 cases) was substantially lower than that reported in the US population-based cohort from the Nurses' Health Study (>14.0 ng/ml = 31.7 nmol/l; 120 cases) in which a higher consumption of folic-acid containing foods and an inverse association between plasma folate levels and BC risk were observed.[9] Thus, a minimal level of blood folate might be required for observing a beneficial effect of this nutrient on the risk of BC.

The high plasma folate concentrations reported in US-based population studies is likely due to folic acid fortification of flour and cereal-grain products, which became mandatory in the United States since 1997 to prevent neural tube defects.[36,37] On the other hand, no policy of folic acid fortification of foods has been implemented in European countries. Folic acid from fortified foods or supplementation is



		Plasma folate†		
		Low	Medium	High
Alcohol intake‡		(<10.96 nmol/l)	(10.96-17.85 nmol/l)	(>17.85 nmol/l)
High (>12 g/day)	OR (95% CI)	1 (ref)	1.14 (0.86; 1.53)	1.02 (0.76; 1.36)
	BC cases	223	229	225
Medium (3-12 g/day)	OR (95% CI)	0.80 (0.57; 1.11)	0.75 (0.54; 1.03)	0.80 (0.58; 1.11)
	BC cases	185	242	231
Low (<3 g/day)	OR (95% CI)	0.74 (0.54; 1.02)	0.86 (0.62; 1.19)	0.75 (0.54; 1.04)
	BC cases	395	383	378
P interaction§				0.69

Figure 1. Multivariable odds ratios (ORs) and 95% confidence intervals (CIs) for association with breast cancer risk by levels of plasma folate (nmol/l) and alcohol intake (g/day), including interaction test. Subjects were matched by study center, age, menopausal status, exogenous hormone use, fasting status and phase of the menstrual cycle. Models were adjusted by date at blood collection, education, BMI, height, physical activity, ever use of hormone replacement therapy, alcohol intake, parity and age at first full-term birth combined, total energy intake and family history of breast cancer. †Tertiles of plasma folate. ‡Categories of alcohol intake (0–3, 3–12 and >12 g/d). §*P* interaction between plasma folate and alcohol intake as categorical variables. All statistical tests were two-sided.

estimated to be approximately 1.7 times more bioavailable than natural folates.[38] Because most of the enzymes that use folate as cofactor cannot use the synthetic form, there might be important perturbations in one-carbon metabolism and cellular processes that rely on this pathway. A recent dose–response meta-analysis of 16 prospective studies including a total of 26,205 BC patients identified a U-shaped relationship between energy-adjusted dietary folate intake and BC risk,[39] supporting prior evidence of an increased risk of BC associated with folic acid fortification.[40] The lack of data on consumption of folic acid-containing supplements within the EPIC population prevented us from testing whether folic acid intake might have been associated with high levels of plasma folate and an increased BC risk. However, the proportion of vitamin supplement users in our study population was only 23% among cases, suggesting that plasma levels of folate and other B vitamins were primarily attributable to natural food sources.

The lack of a significant interaction between plasma folate levels and alcohol intake on BC risk in our analysis is consistent with results from previous prospective studies that used biomarkers of folate status.[8–12] However, a recent prospective investigation within the EPIC study reported an inverse association between dietary folate intake and the risk of BC among heavy alcohol drinkers.[17] Since alcohol may impair folate absorption,[34,39] alcohol consumption behaviors are more likely to modify the risk of BC associated with dietary folate intake rather than plasma folate levels, which can be affected by a variety of other factors including genetic polymorphisms.[41] Thus, women with high intake of both folate and alcohol may not necessarily have a high folate status and consequently a reduced risk of BC.

The main sources of vitamin B12 are animal products, including meat, fish, dairy products, eggs and liver. Our finding of a positive association between plasma levels of vitamin B12 and BC risk in subgroup analyses is in accordance with

Table 4. Crude and multivariable odds ratios[1] for association of plasma folate and vitamin B12 with breast cancer risk stratified by levels of alcohol intake and plasma folate

Plasma vitamin B12 (pmol/l)	Cases/ controls (n)[2]	Crude OR	Multivariable OR[1] (95% CI)	P_{trend}[3]	$P_{interaction}$[4]
Alcohol intake at recruitment[5]					0.14
Below median value (<3.36 g/day)					
Continuous[6]	1,205/1,266	0.99	1.06 (0.87, 1.29)		
<293.6	301/297	1 (ref)	1 (ref)	0.56	
293.6–373.1	296/307	0.95	0.96 (0.76, 1.21)		
373.1–460.0	264/319	0.82	0.85 (0.67, 1.07)		
>460.0	344/343	0.99	1.08 (0.86, 1.35)		
Above median value (≥3.36 g/day)					
Continuous[6]	1,284/1,255	1.2	1.21 (0.97, 1.51)		
<293.6	312/334	1 (ref)	1 (ref)	0.05	
293.6–373.1	332/323	1.1	1.09 (0.87, 1.37)		
373.1–460.0	312/311	1.07	1.07 (0.86, 1.35)		
>460.0	328/287	1.22	1.26 (1.00, 1.58)		
Plasma folate at blood collection[5]					0.05
Below median value (<13.56 nmol/l)					
Continuous[6]	1,218/1,261	1.2	1.25 (1.02, 1.52)		
<293.6	322/377	1 (ref)	1 (ref)	0.03	
293.6–373.1	329/334	1.15	1.15 (0.92, 1.42)		
373.1–460.0	282/278	1.19	1.22 (0.97, 1.53)		
>460.0	285/272	1.23	1.29 (1.02, 1.62)		
Above median level (≥13.56 nmol/l)					
Continuous[6]	1,271/1,260	0.93	0.99 (0.79, 1.22)		
<293.6	291/254	1 (ref)	1 (ref)	0.68	
293.6–373.1	299/296	0.88	0.90 (0.71, 1.13)		
373.1–460.0	294/352	0.73	0.75 (0.59, 0.95)		
>460.0	387/358	0.94	1.02 (0.81, 1.28)		

[1]Models were adjusted by matching factors (study center, age, menopausal status, exogenous hormone use, fasting status and phase of the menstrual cycle), education, BMI, height, physical activity, ever use of hormone replacement therapy, alcohol intake, parity and age at first full-term birth combined, total energy intake and family history of breast cancer.
[2]Cutpoints of quartiles determined on control individuals.
[3]Obtained by modeling the median value of tertiles as continuous term in the multivariable model.
[4]Obtained by modeling the interaction term between plasma vitamin B12 in continuous and alcohol intake or plasma folate as dichotomous variable.
[5]Alcohol intake and plasma folate dichotomized according to median value.
[6]The OR (95% CI) in the continuous model corresponds to an increment of 2.7 units of folate (nmol/l) or vitamin B12 (pmol/l).

two previous prospective studies that measured either dietary intake[42] or plasma levels[11] of this nutrient. However, an inverse association between biomarkers of vitamin B12 and the risk of BC has also been reported.[9,10] The median value of plasma vitamin B12 in our study population (377 = 511 pg/ml in cases) was not substantially different from those reported in other population-based prospective studies, ranging between 421 and 467 pg/ml.[9–11] Thus, several other factors might have contributed to the inconsistent findings, including differences in alcohol consumption, genetic polymorphisms, and nutrient interactions in one-carbon metabolism.[34,43]

As a cofactor required for the generation of methyl groups, a high vitamin B12 status could result in hypermethylation of CpG island promoters for tumor suppressor genes,[44] which may lead to reduced expression of these cancer-related genes and ultimately promote breast carcinogenesis.[45] These DNA methylation changes may also impair the proper expression and/or function of cell-cycle regulatory genes and thus confer a selective growth advantage to neoplastic cells.[46] A randomized crossover trial suggested that moderate alcohol intake may diminish plasma vitamin B12 concentrations.[47] In contrast, a case–control study found that plasma levels of vitamin B12 in heavy alcohol drinkers were

significantly higher than those in light alcohol drinkers.[48] Further studies are needed to clarify the modifying effect of alcohol on the association between vitamin B12 and BC risk.

The positive association between plasma levels of vitamin B12 and BC risk among women with low folate status is unexpected. Previous prospective studies found no evidence of an interaction between these two nutrients on the risk of BC.[16,42,49,50] On the other hand, a prospective analysis within the French E3N cohort reported a strong joint protective effect of high intake of folate and vitamin B12 on BC risk.[18] The almost exclusive form of folate in plasma is 5-methyl THF, which reflects the amount of folate available for DNA methylation.[51] 5-methyl THF is converted to tetrahydrofolate (THF) via the vitamin B12-dependent enzyme methionine synthase. A high vitamin B12 status indicates that methionine synthase activity is increased, leading to depletion of 5-methyl THF and thus plasma folate concentration if not replaced by new 5-methyl THF from diet. In this situation, cells lack the substrate needed for methionine synthesis and DNA methylation is impaired. There is evidence that a low folate status may induce carcinogenesis through alteration of DNA methylation pathways.[52] Thus, the possibility that low plasma folate concentrations (mainly 5-methyl THF) as a consequence of high vitamin B12 status would impair DNA methylation might be suggested.

Epidemiological studies provide support that the association between the *MTHFR* C677T polymorphism and BC risk is modified by intakes of some B vitamins, including folate and vitamin B12.[53–55] We observed no significant effect modification of *MTHFR* SNPs on the association between plasma folate or vitamin B12 and BC risk. The low power of these subgroup analyses prevented us from finding a potential interaction between *MTHFR* genotypes and B vitamin status on the risk of BC. Furthermore, the effect of *MTHFR* polymorphisms on plasma levels of B vitamins is highly complex and may depend on the interaction with other dietary and genetic factors.[56]

The present study is the largest prospective investigation to date to have examined the association between biomarkers of folate and vitamin B12 and the risk of BC. The high follow-up rates and large number of cases provided sufficient statistical power for most subgroup analyses. The major strength of our study is, however, the collection of blood samples prior to diagnosis and the use as biomarkers of exposure as reflection of true vitamin status.

Major limitations include the single collection of blood samples at baseline and the measurement of a single biomarker of folate or vitamin B12 status. Folate concentration measured in plasma is considered to be a sensitive biomarker of recent dietary intake, and thus is not very informative for the assessment of long-term folate status.[57] Plasma vitamin B12 is the most widely used biomarker of total cobalamin status, but not the most specific biomarker to characterize adequate vitamin concentrations.[58] In order to obtain more reliable information on vitamin status, multiple measurements of plasma biomarkers should be taken over a period of time or a combination of different biomarkers should be used. Additional limitations include (*i*) the large percentage of missing data for family history of BC (53.1%) and supplement use (54.5%), (*ii*) the determination of menopausal status at recruitment and not at diagnosis, (*iii*) the lack of complete hormone receptor status data and (*iv*) the insufficient statistical power for gene-nutrient interaction analyses. Because controlling for family history of BC and supplement use had minimal effect on the risk estimates, our results are unlikely to be explained by residual confounding by those factors.

In conclusion, no clear support for an association between plasma levels of folate and BC risk was found in this large prospective study. However, potential interactions between vitamin B12 and alcohol or folate on the risk of BC were observed. Our findings suggest a potential role of vitamin B12 in breast carcinogenesis and raise the possibility of important nutrient–nutrient and gene–nutrient interactions in the etiology of BC. The potential deleterious effect of high vitamin B12 status in combination with other risk factors for BC deserves further investigation. Given the inconsistent findings to date and the possibility that associations between folate and BC could be influenced by some factors yet to be identified, further studies based on novel biomarkers that take into account the effect of potential risk factors and genetic polymorphisms are warranted.

Acknowledgements

The authors gratefully acknowledge Deborah Postoly, laboratory technician, for sample processing and genotyping of study samples, as well as chief engineer Gry Kvalheim within the Bevital laboratory, Bergen, Norway, for her outstanding assistance with laboratory measurements of blood B vitamins. We confirm that all authors listed have contributed to the planning, execution and analysis of the submitted manuscript and that there are no conflicts of interest whatsoever.

References

1. Hankinson SE, Colditz GA, Willett WC. Towards an integrated model for breast cancer etiology: the lifelong interplay of genes, lifestyle, and hormones. *Breast Cancer Res* 2004;6:213–8.
2. Breast Cancer 2010 Report. Food, nutrition, physical activity, and the prevention of breast cancer. 2016.
3. Ames BN, Wakimoto P. Are vitamin and mineral deficiencies a major cancer risk?. *Nat Rev Cancer* 2002;2:694–704.
4. Vera-Ramirez L, Ramirez-Tortosa MC, Sanchez-Rovira P, et al. Impact of diet on breast cancer risk: a review of experimental and observational studies. *Crit Rev Food Sci Nutr* 2013;53:49–75.
5. Mason JB. Biomarkers of nutrient exposure and status in one-carbon (methyl) metabolism. *J Nutr* 2003;133(Suppl.3):941S–7S.
6. Davis CD, Uthus EO. DNA methylation, cancer susceptibility, and nutrient interactions. *Exp Biol Med (Maywood)* 2004;229:988–95.
7. Szyf M, Pakneshan P, Rabbani SA. DNA methylation and breast cancer. *Biochem Pharmacol* 2004;68:1187–97.
8. Rossi E, Hung J, Beilby JP, et al. Folate levels and cancer morbidity and mortality: prospective cohort study from Busselton, Western Australia. *Ann Epidemiol* 2006;16:206–12.
9. Zhang SM, Willett WC, Selhub J, et al. Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer. *J Natl Cancer Inst* 2003;95:373–80.

10. Wu K, Helzlsouer KJ, Comstock GW, et al. A prospective study on folate, B12, and pyridoxal 5′-phosphate (B6) and breast cancer. *Cancer Epidemiol Biomarkers Prev* 1999;8:209–17.
11. Lin J, Lee IM, Cook NR, et al. Plasma folate, vitamin B-6, vitamin B-12, and risk of breast cancer in women. *Am J Clin Nutr* 2008;87:734–43.
12. Ericson UC, Ivarsson MI, Sonestedt E, et al. Increased breast cancer risk at high plasma folate concentrations among women with the MTHFR 677T allele. *Am J Clin Nutr* 2009;90:1380–9.
13. Wu W, Kang S, Zhang D. Association of vitamin B6, vitamin B12 and methionine with risk of breast cancer: a dose-response meta-analysis. *Br J Cancer* 2013;109:1926–44.
14. Ericson U, Sonestedt E, Gullberg B, et al. High folate intake is associated with lower breast cancer incidence in postmenopausal women in the Malmo Diet and Cancer cohort. *Am J Clin Nutr* 2007;86:434–43.
15. Stolzenberg-Solomon RZ, Chang SC, Leitzmann MF, et al. Folate intake, alcohol use, and postmenopausal breast cancer risk in the prostate, lung, colorectal, and ovarian cancer screening trial. *Am J Clin Nutr* 2006;83:895–904.
16. Stevens VL, McCullough ML, Sun J, et al. Folate and other one-carbon metabolism-related nutrients and risk of postmenopausal breast cancer in the Cancer Prevention Study II Nutrition Cohort. *Am J Clin Nutr* 2010;91:1708–15.
17. de BJ, Ferrari P, Chajes V, et al. Dietary folate intake and breast cancer risk: European prospective investigation into cancer and nutrition. *J Natl Cancer Inst* 2015;107:dju367.
18. Lajous M, Romieu I, Sabia S, et al. Folate, vitamin B12 and postmenopausal breast cancer in a prospective study of French women. *Cancer Causes Control* 2006;17:1209–13.
19. Choi SW, Mason JB. Folate and carcinogenesis: an integrated scheme. *J Nutr* 2000;130:129–32.
20. Koushik A, Kraft P, Fuchs CS, et al. Nonsynonymous polymorphisms in genes in the one-carbon metabolism pathway and associations with colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2006;15:2408–17.
21. Sharp L, Little J. Polymorphisms in genes involved in folate metabolism and colorectal neoplasia: a HuGE review. *Am J Epidemiol* 2004;159:423–43.
22. Hazra A, Wu K, Kraft P, et al. Twenty-four nonsynonymous polymorphisms in the one-carbon metabolic pathway and risk of colorectal adenoma in the Nurses' health study. *Carcinogenesis* 2007;28:1510–9.
23. Frosst P, Blom HJ, Milos R, et al. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat Genet* 1995;10:111–3.
24. van der Put NM, Gabreels F, Stevens EM, et al. A second common mutation in the methylenetrahydrofolate reductase gene: an additional risk factor for neural-tube defects?. *Am J Hum Genet* 1998;62:1044–51.
25. Bingham S, Riboli E. Diet and cancer – the European Prospective Investigation into Cancer and Nutrition. *Nat Rev Cancer* 2004;4:206–15.
26. Riboli E, Hunt KJ, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 2002;5:1113–24.
27. 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–S14.
28. Slimani N, Deharveng G, Unwin I, et al. The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr* 2007;61:1037–56.
29. Bouckaert KP, Slimani N, Nicolas G, et al. Critical evaluation of folate data in European and international databases: recommendations for standardization in international nutritional studies. *Mol Nutr Food Res* 2011;55:166–80.
30. Ritte R, Tikk K, Lukanova A, et al. Reproductive factors and risk of hormone receptor positive and negative breast cancer: a cohort study. *BMC Cancer* 2013;13:584.
31. Molloy AM, Scott JM. Microbiological assay for serum, plasma, and red cell folate using cryopreserved, microtiter plate method. *Methods Enzymol* 1997;281:43–53.
32. Kelleher BP, Broin SD. Microbiological assay for vitamin B12 performed in 96-well microtitre plates. *J Clin Pathol* 1991;44:592–5.
33. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med* 1989;8:551–61.
34. Halsted CH, Villanueva JA, Devlin AM, et al. Metabolic interactions of alcohol and folate. *J Nutr* 2002;132(Suppl. 8):2367S–72S.
35. Leenders M, Ros MM, Sluijs I, et al. Reliability of selected antioxidants and compounds involved in one-carbon metabolism in two Dutch cohorts. *Nutr Cancer* 2013;65:17–24.
36. Pfeiffer CM, Johnson CL, Jain RB, et al. Trends in blood folate and vitamin B-12 concentrations in the United States, 1988 2004. *Am J Clin Nutr* 2007;86:718–27.
37. MMWR. Trends in wheat-flour fortification with folic acid and iron–worldwide, 2004 and 2007. *MMWR Morb Mortal Wkly Rep* 2008;57:8–10.
38. Caudill MA. Folate bioavailability: implications for establishing dietary recommendations and optimizing status. *Am J Clin Nutr* 2010;91:1455S–60S.
39. Chen P, Li C, Li X, et al. Higher dietary folate intake reduces the breast cancer risk: a systematic review and meta-analysis. *Br J Cancer* 2014;110:2327–38.
40. Kim YI. Does a high folate intake increase the risk of breast cancer?. *Nutr Rev* 2006;64(10 Part 1):468–75.
41. Lalouschek W, Aull S, Serles W, et al. Genetic and nongenetic factors influencing plasma homocysteine levels in patients with ischemic cerebrovascular disease and in healthy control subjects. *J Lab Clin Med* 1999;133:575–82.
42. Bassett JK, Baglietto L, Hodge AM, et al. Dietary intake of B vitamins and methionine and breast cancer risk. *Cancer Causes Control* 2013;24:1555–63.
43. Powers HJ. Interaction among folate, riboflavin, genotype, and cancer, with reference to colorectal and cervical cancer. *J Nutr* 2005;135(Suppl. 12):2960S–6S.
44. Baylin SB, Herman JG, Graff JR, et al. Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res* 1998;72:141–96.
45. Widschwendter M, Jones PA. DNA methylation and breast carcinogenesis. *Oncogene* 2002;21:5462–82.
46. Gonzalgo ML, Jones PA. Mutagenic and epigenetic effects of DNA methylation. *Mutat Res* 1997;386:107–18.
47. Laufer EM, Hartman TJ, Baer DJ, et al. Effects of moderate alcohol consumption on folate and vitamin B(12) status in postmenopausal women. *Eur J Clin Nutr* 2004;58:1518–24.
48. Cyłwik B, Czygier M, Daniluk M, et al. Vitamin B12 concentration in the blood of alcoholics. *Pol Merkur Lekarski* 2010;28:122–5.
49. Kabat GC, Miller AB, Jain M, et al. Dietary intake of selected B vitamins in relation to risk of major cancers in women. *Br J Cancer* 2008;99:816–21.
50. Cho E, Holmes M, Hankinson SE, et al. Nutrients involved in one-carbon metabolism and risk of breast cancer among premenopausal women. *Cancer Epidemiol Biomarkers Prev* 2007;16:2787–90.
51. Fazili Z, Pfeiffer CM. Measurement of folates in serum and conventionally prepared whole blood lysates: application of an automated 96-well plate isotope-dilution tandem mass spectrometry method. *Clin Chem* 2004;50:2378–81.
52. Crider KS, Yang TP, Berry RJ, et al. Folate and DNA methylation: a review of molecular mechanisms and the evidence for folate's role. *Adv Nutr* 2012;3:21–38.
53. Chen J, Gammon MD, Chan W, et al. One-carbon metabolism, MTHFR polymorphisms, and risk of breast cancer. *Cancer Res* 2005;65:1606–14.
54. Shrubsole MJ, Gao YT, Cai Q, et al. MTHFR polymorphisms, dietary folate intake, and breast cancer risk: results from the Shanghai Breast Cancer Study. *Cancer Epidemiol Biomarkers Prev* 2004;13:190–6.
55. Maruti SS, Ulrich CM, Jupe ER, et al. MTHFR C677T and postmenopausal breast cancer risk by intakes of one-carbon metabolism nutrients: a nested case-control study. *Breast Cancer Res* 2009;11:R91.
56. Coughlin SS, Piper M. Genetic polymorphisms and risk of breast cancer. *Cancer Epidemiol Biomarkers Prev* 1999;8:1023–32.
57. Green R. Indicators for assessing folate and vitamin B-12 status and for monitoring the efficacy of intervention strategies. *Am J Clin Nutr* 2011;94:666S–72S.
58. EFSA Panel on Dietetic Products NaAN. Scientific opinion on dietary reference values for cobalamin (vitamin B12). *EFSA J* 2015;13:4150.

2- Dietary folate, alcohol consumption and DNA methylation

Context

The one-carbon metabolism (OCM) is a network of interrelated biochemical reaction in which a one-carbon unit is received from methyl donor nutrients and transferred into biochemical and molecular pathways essential for DNA replication and repair. Modifications in OCM can significantly impact gene expression and thereby cellular function (65). There is increasing evidence that folate, as one of the methyl donor nutrients, is a relevant candidate for modulation of the epigenome (83). Alcohol metabolites, involved in a dysfunction of the folate absorption, have also shown to affect the epigenome. This antagonist effect of alcohol on folate could plausibly increase the need of folate intake. Inadequate folate level may result in abnormal DNA synthesis and disrupted DNA repair and hence may influence cancer risk, including breast cancer (69). However the epidemiological evidence linking dietary folate, alcohol intake and epigenome modifications is not well documented.

Objectives

- To identify single CpG sites differentially methylated in relation to dietary folate and alcohol intake.
- To investigated the association between dietary folate and alcohol intake with DNA methylation levels in regions of CpG sites.

Approach

Genome-wide DNA profiles on about 450,000 CpG sites were measured using Illumina Infinium HumanMethylation450K in 450 cancer-free women, part of a nested case-control study on BC within the EPIC cohort. SVA normalization technique was used to remove unwanted variation from DNA methylation introduced by samples processing during methylation acquisition such as the batch. Dietary folate and alcohol intake were assessed at recruitment through questionnaires.

In this study the association of dietary folate and alcohol intake with DNA methylation was investigated via three different approaches. The site-specific analysis aimed at identifying single CpG site independently from each other, whereas Differentially Methylated Regions (DMRs) analysis (20) and fused lasso (FL) regressions (84) analyses aimed at identifying regions of CpG sites. The latter approaches use the hypothesis that neighboring CpG sites may share similar information, thus exploiting the potential of specific regions of the epigenome to show methylation activity related to lifestyle factors. FDR was used to control statistical tests for multiple testing.

Main findings

After correction for multiple testing, site-specific analysis showed a lack of association between dietary folate and individual CpG sites. Alcohol intake was positively associated with methylation level in cg03199996, and inversely associated with methylation in cg07382687. These two associations were borderline significant (both q_{val}=0.049). A total of 24 and 90 differentially methylated regions (DMRs) were associated with dietary folate and alcohol intake, respectively. An inverse association was observed for 54% of the dietary folate DMRs and for 44% of the alcohol intake DMRs. FL regression identified 71 regions significant for dietary folate including 70% with an inverse association and 133 regions significant for alcohol intake including 47% with an inverse association. However, the overlap between the two methods was relatively low, i.e. three and 21 FL regions were overlapping dietary folate and alcohol intake DMRs, respectively. There was an especially high concentration of regions in chromosome 6 where 4 DMRs were overlapping FL regions and in chromosome 22 counting 3 overlaps between the DMRs and FL regions.

Conclusion

A borderline association between alcohol intake and methylation levels in two CpG sites was observed. Evidence from DMRs an FL analysis indicated that both dietary folate and alcohol intake might be associated with alteration of DNA methylation levels in localized regions. Folate and alcohol are suspected to be associated with breast cancer risk but also to have antagonist roles in the one-carbon metabolism. In certain regions identified by DMRs or FL analysis, mapped genes are known to act as tumor suppressor such as the *GSDMD and HOXA5* genes. These results were in line with the hypothesis that folate- and alcohol-deregulated epigenetic mechanisms might have a role in the pathogenesis of cancer.

Submitted article: Association of leukocyte DNA methylation changes with dietary folate and alcohol intake in the EPIC Study.

The following draft has been recently submitted and is under consideration at Clinical Epigenetics.

Association of leukocyte DNA methylation changes with dietary folate and alcohol intake in the EPIC Study

Perrier F[1], Viallon V[1], Ambatipudi S[2,3], Ghantous A[2], Cyrille Cuenin[2], Hernandez-Vargas H[2], Chajès V[4], Baglietto L[5], Matejcic M[4,6], Moreno-Macias H[7], Kühn T[8], Boeing H[9], Karakatsani A[10,11], Kotanidou A[10,12], Trichopoulou A[10], Sieri S[13], Panico S[14], Fasanelli F[15], Dolle M[16], Onland-Moret C[17], Sluijs I[17], Weiderpass E[18,19,20,21], Quirós JR[22], Agudo A[23], Huerta JM[24,25], Ardanaz E[24,26], Dorronsoro M[27], Tong TYN[28], Tsilidis K[29], Riboli E[29], Gunter M.J[4], Herceg Z[2], Ferrari P[1,#,*] and Romieu I[4,#].

[1]Nutritional Methodology and Biostatistics Group, International Agency for Research on Cancer (IARC), Lyon, France;

[2]Epigenetics Group, IARC, Lyon, France;

[3]MRC integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK;

[4]Nutritional Epidemiology Group, IARC, Lyon, France;

[5]Department of Clinical and Experimental Medicine, University of Pisa, Italy;

[6]Department of Preventive Medicine, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, USA;

[7]Universidad Autonoma Metropolitana, Mexico City, Mexico;

[8]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany;

[9]Department of Epidemiology, German Institute of Human Nutrition (DIfE), Potsdam-Rehbrücke, Germany;

[10]Hellenic Health Foundation, Athens, Greece;

[11]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Greece;

[12]1st Department of Critical Care Medicine & Pulmonary Services, University of Athens Medical School, Evangelismos Hospital, Athens, Greece;

[13]Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy;

[14]Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy;

[15]Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Via Santena 7, Turin, Italy;

[16]National Institute of Public Health and the Environment (RIVM),Centre for Health Protection (pb12), Bilthoven, Netherlands;
[17]Department of Epidemiology, Julius Center Research Program Cardiovascular Epidemiology, Utrecht, The Netherlands;
[18]Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway;
[19]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden;
[20]Genetic Epidemiology Group, Folkhälsan Research Center and Faculty of Medicine, University of Helsinki, Helsinki, Finland;
[21]Department of Community Medicine, University of Tromsø, The Arctic University of Norway, Tromsø, Norway;
[22]Public Health Directorate, Asturias, Spain;
[23]Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain;
[24]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain;
[25]CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain;
[26]Navarra Public Health Institute, Pamplona, Spain & IdiSNA, Navarra Institute for Health Research, Pamplona, Spain & CIBER Epidemiology and Public Health CIBERESP, Madrid, Spain;
[27] Public Health Direction and Biodonostia Research Institute and Ciberesp, Basque Regional Health Department, San Sebastian, Spain;
[28]Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford;
[29]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK.
[#]Joint Senior Authors.

*Corresponding author:
Pietro Ferrari, PhD
Nutritional Methodology and Biostatistics Group (NMB)
International Agency for Research on Cancer
World Health Organization

150, cours Albert Thomas
69372 Lyon CEDEX 08, France
Off. +33 472 738 031
Mob. + 33 668 728146
E-mail: ferrarip@iarc.fr

Abstract (Words=330)

Background: There is increasing evidence that folate, an important component of one-carbon metabolism, modulates the epigenome. Alcohol, which can disrupt folate absorption, is also known to affect the epigenome. We investigated the association of dietary folate and alcohol intake on leukocyte DNA methylation levels in the European Prospective Investigation into Cancer and nutrition (EPIC) study. Leukocyte genome-wide DNA methylation profiles on approximately 450,000 CpG sites were acquired with Illumina HumanMethylation 450K BeadChip measured among 450 women control participants of a case-control study on breast cancer nested within the EPIC cohort. After data pre-processing using Surrogate Variables Analysis to reduce systematic variation, associations of DNA methylation with dietary folate and alcohol intake, assessed with dietary questionnaires, were investigated using CpG site-specific linear models. Specific regions of the methylome were explored using Differentially Methylated Regions (DMR) analysis and fused lasso (FL) regressions. The DMRs analysis combined results from feature-specific analysis for a specific chromosome and using distances between features as weights whereas FL regression combined two penalties to encourage sparsity of single features and the difference between two consecutive features.

Results: After correction for multiple testing, intake of dietary folate was not associated with methylation level at any DNA methylation site, while alcohol intake was positively associated with methylation level at cg03199996 (q_{val}=0.049), and inversely associated with methylation level at cg07382687 (q_{val}=0.049). Interestingly, the DMR analysis revealed a total of 24 and 90 regions associated with dietary folate and alcohol, respectively. For alcohol intake, 6 of the 15 most significant DMRs were identified through FL.

Conclusions: Alcohol intake was associated with methylation levels at two CpG sites. Evidence from DMRs and FL analysis indicated that dietary folate and alcohol intake may be associated with genomic regions with tumor suppressor activity such as the *GSDMD* and *HOXA5* genes. These results were in line with the hypothesis that epigenetic mechanisms play a role in the association between folate and alcohol, although further studies are warranted to clarify the importance of these mechanisms in cancer.

Keywords: DNA methylation, dietary folate, alcohol intake, DMR, fused lasso, EPIC cohort.

Introduction

DNA methylation is a crucial epigenetic mechanism involved in regulating important cellular processes, including gene expression, cell differentiation, genomic imprinting and preservation of chromosome stability. DNA methylation refers to the addition of methyl groups (-CH3) to the carbon-5 position of cytosine residues in a cytosine-guanine DNA sequence (CpG) by DNA methyltransferases. DNA methylation changes can be influenced by many factors including aging (Heyn et al., 2012; Horvath & Raj, 2018) and environmental exposure such as smoking (S. Ambatipudi et al., 2016; Joehanes et al., 2016) or specific dietary factors (Niculescu & Haggarty, 2011). Experimental evidence suggests a link between B-vitamins, including folate (vitamin B9), and epigenetics modifications (Ba et al., 2011). B-vitamins, especially folate, are essential components of one-carbon metabolism (OCM), the network of interrelated biochemical reaction in which a one-carbon unit is received from methyl donor nutrients and transferred into biochemical and molecular pathways essential for DNA replication and repair. Modifications in OCM can significantly impact gene expression and thereby cellular function (Szyf, 2011).

Absorbed folate circulating in the bloodstream enters the OCM cycle in the liver where is metabolize to 5-methyltetrahydrofolate (5-methylTHF) and converted into S-adenosylmethyonine (SAM) after several successive transformation steps (Figure 1). SAM is the methyl donor for numerus methylation reactions including the methylation of DNA, RNA and proteins. The potential role of specific dietary factors including micronutrients such as folate, alcohol, and soya intake, in modifying breast cancer risk via epigenetic mechanisms has been proposed (Teegarden, Romieu, & Lelievre, 2012), although evidence is still scarce and inconsistent.

Alcohol intake affects epigenetic profiles (Liu et al., 2016). Ethanol metabolism generates toxins that may directly lead to OCM dysfunction by reducing folate absorption, increasing renal excretion of folate and inhibiting methionine synthase, the key enzyme in the generation of the methyl donor in the OCM (Liu et al., 2016; Mason & Choi, 2005). This antagonistic effect of alcohol on folate could plausibly increase the need of folate intake. Inadequate folate levels may result in abnormal DNA synthesis due to a reduced availability of SAM (Kruman & Fowler, 2014) and disrupted DNA repair and may, hence, influence cancer risk, including breast cancer (Baglietto, English, Gertig, Hopper, & Giles, 2005; Zhang et al., 1999).

135 The epidemiological evidence linking dietary folate, alcohol intake and epigenome
136 modifications is, however, not well documented. Therefore, we investigated the relationships
137 between dietary folate and alcohol intake with leukocyte DNA methylation patterns in the
138 controls from the European Prospective Investigation into Cancer and nutrition (EPIC) study
139 on breast cancer. We complemented standard regression analysis with techniques for the
140 identification of relevant methylated regions.

141

142 **Methods**

143 *Study population*. EPIC is a multicentre study that recruited over 521,000 participants,
144 between 1992 and 2000 in 23 regional or national centres in 10 European countries (Denmark,
145 France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and United Kingdom)
146 (Riboli et al., 2002). Among the 367,903 women recruited in EPIC, and after exclusion of
147 19,583 participants with prevalent cancers at recruitment (except non-melanoma skin cancer),
148 first malignant primary BC occurred for 10,713 women during follow-up between 1992 and
149 2010. Within a nested case-control study that included 2,491 invasive BC cases (Matejcic et
150 al., 2017), a subsample of 960 women (480 cases and 480 matched controls) from Germany,
151 Greece, Italy, Netherlands, Spain and United Kingdom was selected for the DNA methylation
152 analyses (Srikant Ambatipudi et al., 2017). The present study included analysis of 450
153 controls only originally enrolled in this case-control study on breast cancer (BC) nested
154 within EPIC study.

155 *Methylation acquisition.* Genome-wide DNA-methylation profiles in buffy coat samples were
156 quantified using the Illumina Infinium HumanMethylation450K (HM450K) BeadChip assay
157 (Bibikova et al., 2011) in 960 biospecimens from women included in the BC nested case-
158 control study. A total of 20 biospecimens with replicates used to compare technical inter- and
159 intra-assay batch effects and then excluded from the main analysis together with 19 matched
160 pairs, i.e. 38 samples, where at least one of the two samples had a low-quality bisulfite
161 conversion efficiency (intensity signal<4000) or did not pass all of the Illumina
162 GenomeStudio quality control steps, which were based on built-in control probes for staining,
163 hybridization, extension, and specificity (Illumina, 2011). To prevent collider bias (Cole et al.,
164 2010), as both alcohol intake and folate intake and DNA-methylation profiles are all
165 potentially associated with causes of BC, among the 902 remaining samples from the original
166 case-control study on BC nested within EPIC study, only cancer-free women were selected

for the present study. For the 451 controls sample, probes with detection p-values higher than 0.05 were assigned 'missing' value. After the exclusion of 14,548 cross-reactive probes (Y. A. Chen et al., 2013), 47,963 probes overlapping known SNPs with minor allele frequency (MAF) greater than 5% in the overall population (European ancestry) (Y. A. Chen et al., 2013) and 1,483 low quality probes (i.e. missing in more than 5% of the samples), 421,583 probes were left for the statistical analyses (Srikant Ambatipudi et al., 2017).

For each probe, β-values were calculated as the ratio of methylated intensity over the overall intensity, defined as the sum of methylated and unmethylated intensities. The following preliminary adjustment steps were applied to β-values: (i) color bias normalization using smooth quantile normalization (P. Du, Kibbe, & Lin, 2008); (ii) quantile normalization (Bolstad, 2001); (iii) type I and type II bias correction using the Beta-Mixture Quantile normalization (BMIQ) (Teschendorff et al., 2013). Then, M-values, defined as $M_{values} = log_2\left(\frac{\beta_{values}}{1-\beta_{values}}\right)$, were computed (Pan Du et al., 2010). Surrogate Variables Analysis (SVA) (J. T. Leek & Storey, 2007, 2008) was used to remove systematic variation due to the processing of the biospecimens during methylation acquisition such as batch, indicating groups of samples processed at the same time, and the position of the samples within the chip (Perrier et al., 2018).

The percentage of white blood cell counts, i.e. T cells ($CD8^+T$ and $CD4^+T$), natural killer (NK) cells, B cells, monocytes and granulocytes, was quantified using Houseman's estimation method (Houseman et al., 2012) and included as covariates in the analysis.

Lifestyle and dietary exposures. Data on dietary habits were collected at recruitment through validated centre- or country-specific dietary questionnaires (DQ) (Riboli et al., 2002). Northern Italy (Florence, Turin and Varese), United Kingdom, Germany and the Netherland used self-administered extensive quantitative food-frequency questionnaires (FFQs), whereas Southern Italy (Naples and Ragusa), Spain and Greece's centres used interview methods. Usual consumption of alcoholic beverages (number of glasses per day or week) per type of alcoholic beverage (wine, beer, spirits and liquors) during the 12 months before the administration of dietary questionnaires was collected at recruitment. Alcohol intake in g/day was calculated combining all types of beverage for each country based on the estimated average of glass volume and ethanol content for each type of beverages (Ferrari et al., 2007; Slimani et al., 2000). Dietary folate intake (µg/day) was estimated using the updated EPIC Nutrient Data Base (ENDB) (Slimani et al., 2007), obtained after standardization from

199 country-specific food composition tables (Bouckaert et al., 2011). No specific information on
200 use of folate supplements was available.

201 *Statistical analyses.* After exclusion of one outlier value of dietary folate (value larger than
202 the third quartile plus 10 times the inter-quartile range of the distribution), a total of 450
203 observations from controls only were retained for statistical analyses.

204 The association between dietary folate, alcohol intake and methylation levels was evaluated
205 via (i) CpG site-specific analysis; (ii) identification of differentially methylated regions
206 (DMRs) (Peters et al., 2015); (iii) fused lasso (FL) regression (Tibshirani, Saunders, Rosset,
207 Zhu, & Knight, 2005).

208 *(i) CpG site-specific models.* M-values expressing methylation levels at each CpG were
209 linearly regressed on dietary folate (log-transformed to reduce skewness) and alcohol intake.
210 Models were adjusted for recruitment centre, age at recruitment (year), menopausal status
211 (pre- or post-menopause) and white blood cell counts (proportions of T cells, natural killer
212 cells, B cells and monocytes in blood). False discovery rate (FDR) was used to control
213 statistical tests for multiple testing.

214 *(ii) DMRs models.* Differentially methylated regions (DMRs) analyses were identified with
215 the *DMRcate* package (Peters et al., 2015). The rationale of this method is to use kernel
216 smoothing to replace the *t*-test statistics at a given CpG site by a weighted average of *t*-test
217 statistics across its neighboring sites on the same chromosome. More precisely, let p_c express
218 the number of sites located on a given chromosome c with $c \in \{1, \ldots, 23\}$ (the 23rd
219 chromosome is chromosome X). For any site k on this chromosome, with $k = 1, \ldots, p_c$, the
220 term ${t_k}^2$ indicates the square of the *t*-test statistics obtained in site-specific analyses. For each
221 site j on chromosome c, ${t_j}^2$ is replaced by the term ${\hat{t}_j}^2$, defined as ${\hat{t}_j}^2 = \sum_{k=1}^{p_c} K_{jk} {t_k}^2$

222 where the terms K_{jk} express weights, with larger values for sites k closer to j. Let x_k express
223 the position of site k on the chromosome, i.e. its chromosomal coordinate in base pairs, these
224 weights are defined using a Gaussian kernel, as

$$K_{jk} = exp\left(\frac{-\left|x_j - x_k\right|^2}{2(\lambda/C)^2}\right)$$

225 where parameters λ and C represent the bandwidth and the scaling factor, respectively. Here
226 we used λ = 1,000 and C = 2, respectively, as recommended in (Peters et al., 2015).

Under the null hypothesis of no association between site j and alcohol (or folate), the distribution of $\frac{\hat{t}_j^{\,2} \sum_k^{p_c} K_{jk}}{\sum_k^n K_{jk}^{\,2}}$ can be approximated by a χ^2 distribution (Peters et al., 2015) with $\left(\sum_k^{p_c} K_{jk}\right)^2 / \sum_k^{p_c} K_{jk}^{\,2}$ degrees of freedom (Satterthwaite, 1946). Accordingly, p-values were obtained for each site separately in each chromosome and q-values were computed using FDR correction on all the p-values to control for multiple-testing. Then, DMRs were defined as regions with at least two significant sites separated by a maximal distance λ of 1000 base pairs. In line with (Peters et al., 2015), *t*-statistics t_k were obtained from regression models using an empirical Bayes method to shrink the CpG site variance (Smyth, 2004), as implemented in the *limma* package (Smyth, 2005). For each DMR, the minimum q-value and the maximum coefficient (in absolute value) of the sites included in the region were presented as q_{DMR} and β_{DMR}.

(iii) Fused lasso regression. Multivariate penalized regression provides an alternative to DMRs. We implemented a Fused Lasso (FL) regression (Tibshirani et al., 2005), which is better suited than the standard lasso when covariates (CpGs) are naturally ordered and the objective is to identify regions on the chromosome of differentially methylated CpG sites. FL is particularly useful when the number of features (p) is way larger than the sample size (n), a situation classically known as $p \gg n$.

FL is a multivariable regression method combining two penalties: (i) the lasso penalty, which introduces sparsity of the parameter vector, i.e. many elements of the estimated vector are encouraged to be set to zero, and (ii) the fused penalty, which encourages sparsity of the difference between two consecutive components in the parameter vector, thus introducing smoothness of parameter estimates in adjacent CpG sites (Tibshirani et al., 2005).

To mimic the DMR analysis, a FL analysis was implemented where dietary folate and alcohol were, in turn, regressed on CpG methylation levels within each chromosome. The vector of methylation coefficient estimates $\hat{\beta}$ obtained by fused lasso regression was defined as

$$\hat{\beta} = \mathbf{arg\,min} \left\{ \sum_i \left(y_i - \sum_j M_{ij}\beta_j - \gamma^T Z_i\right)^2 + \hat{\lambda}_1 \sum_{j=1}^{p_c} \omega_j |\beta_j| + \hat{\lambda}_2 \sum_{j=2}^{p_c} \nu_j |\beta_j - \beta_{j-1}| \right\},$$

where y_i indicate, in turn, alcohol and dietary folate values for sample $i = 1, \ldots, n$, M_{ij} is the methylation levels at CpG site j, β_j is the associated regression coefficient, Z_i is a vector of confounding factors, consistently with linear regression and DMR analyses described above,

γ is the corresponding non-penalized vector of coefficients, ω_j and ν_j are the weights associated with lasso penalty and fused penalty, respectively.

Following the rationale of the adaptive lasso (Zou, 2006) and the iterated lasso (Candès, Wakin, & Boyd, 2008), the FL procedure was run a first time with weights ω_j and ν_j set to 1, which returned $\hat{\beta}_0$, an initial estimate of $\hat{\beta}$. The final estimates $\hat{\beta}$ were obtained after running a second FL procedure with weights defined as $\omega_j = \frac{1}{|\hat{\beta}_{0,j}|+\varepsilon}$ and $\nu_j = \frac{1}{|\hat{\beta}_{0,j}-\hat{\beta}_{0,j-1}|+\varepsilon}$, with $\varepsilon = 10^{-4}$.

The FL procedure was implemented on a predefined grid of 50x50=2,500 values for the pair of parameters (λ_1, λ_2). More precisely, the grid for λ_1 consisted of 50 equally spaced values (on a log scale) between $\frac{\lambda_{1,max}}{1000}$ and $\lambda_{1,max}$, where $\lambda_{1,max}$ was the lowest λ_1 value for which FL returned a null $\hat{\beta}$ vector for λ_2=0, a situation where FL reduces to a standard lasso. For each value $\lambda_1$on this grid, the grid for $\lambda_2$consisted of 50 equally spaced values (on a log-scale) between $\frac{\lambda_{2,max}(\lambda_1)}{1000}$ and $\lambda_{2,max}(\lambda_1)$, where $\lambda_{2,max}(\lambda_1)$ was the lowest λ_2 value for which FL returned a vector $\hat{\beta}$ with all components equal. The optimal pair of tuning parameters (λ_1, λ_2) was selected as the one minimizing the prediction error estimated by 5-fold cross-validation (Hastie, 2009), whose principle can be summarized as follows. The original sample is first partitioned into 5 equally sized subsamples. One subsample is held as the test set while the other 4 are used as a training set, on which FL estimates are computed for the 2,500 values for (λ_1, λ_2). The prediction error is computed on the test set, and the process is repeated 5 times, and for each of the 2,500 values of (λ_1, λ_2). The prediction error is defined as the averaged prediction error on the 5 test sets. FL analysis was implemented using the *FusedLasso* package.

Preprocessing steps and statistical analyses were carried out using the R software (https://www.r-project.org/) and the Bioconductor packages (Huber et al., 2015), including *lumi*, *wateRmelon* and *sva* (Jeffrey T. Leek, Johnson, Parker, Jaffe, & Storey, 2012) for the preprocessing steps. The nominal level of statistical significance was set to 5%.

Results

Study population characteristics. Detailed characteristics of the 450 women included in the study are shown in Table 1. The average age at blood collection was 52 years (range: 26-73).

Participants had an average body mass index (BMI) of 26 kg/m^2 (range: 16-43), and were mostly post-menopausal (59%), never-smokers (56%) and moderately physically inactive (42%). The average daily intakes of dietary folate was 270 µg/day (range: 91-1012) and alcohol daily intake was 8 g/day (range: 0-72). Non-alcohol consumers, defined as participants consuming less than 0.1g/day of alcohol at recruitment, represented 15% of the population. Most participants were from the Italian and the German EPIC centres (Additional file 1, Figure S1).

CpG site-specific models. After FDR correction, dietary folate intake was not significantly associated with methylation levels at any CpG sites (data not shown). Alcohol intake was significantly inversely associated with the cg07382687 CpG site (q_{val}=0.049) and positively associated with the cg03199996 site (q_{val}=0.049) (Table 2). Both sites were located in an open sea region, i.e. a genomic region of isolated CpGs. cg07382687 was within the body region of gene *CREB3L2,* and cg03199996 was within the body region of gene *FAM65C*.

DMRs analysis. A total of 24 regions associated with dietary folate were identified, which included 190 CpG sites over-represented in the TSS1500 and 1st exon regions and under-represented in the body regions and regions outside any gene regions (Figure 2A). The 15 most significant regions are described in Table 3**Error! Reference source not found.** and the whole list provided in Additional file 2, Table S1. Among the 24 DMRs, 54% showed an inverse association with dietary folate, i.e had a $\beta_{\mathrm{DMR}} < 0$. The DMR most significantly associated with dietary folate (q_{DMR}=1.3 E-13, β_{DMR}=0.054), was DMR.F1 in chromosome 7, including 49 CpG sites, related to *HOXA5* and *HOXA6* genes. DMR.F5, was associated with *HOXA4*, another gene of the homeobox family, (q_{DMR}=5.8 E-4, β_{DMR}=−0.047).

Alcohol intake was associated with methylation levels in 90 DMRs, including 550 CpG sites over-represented in TSS200, 1st exon and 5' untranslated regions (5'UTR) and under-represented in the body regions and the regions outside any gene regions (Figure 2B). The 15 most significant DMRs are detailed in Table 4 and the full list is described in Additional file 3, Table S2. Alcohol intake was positively associated with methylation levels in 66% of the 90 DMRs. The two sites associated with alcohol intake in the CpG site-specific analyses were not included in any DMRs. The most significant DMR associated with alcohol consumption was DMR.A1, 9 sites within the *GSDMD* gene, (q_{DMR}=4.7 E-14, β_{DMR}=0.0017).

Methylation levels of each CpG site located in the DMR.A1, DMR.A2, DMR.F1 and DMR.F2, the two most significant DMRs for folate and alcohol, are presented in Additional

file 4, Figure S2 by tertiles of dietary folate and alcohol intake, respectively. Correlation heatmaps of CpG sites in DMR.A1, DMR.A2, DMR.F1 and DMR.F2 are displayed in Additional file 5, Figure S3, showing high levels of correlation among methylation levels within the DMR.F2 of dietary folate and the DMR.A2 of alcohol. Other regions showed less correlation, including the DMRA1 of alcohol intake.

Fused lasso regression. For dietary folate, we identified 71 FL regions, 50 presenting a positive association and 21 an inverse association. Three FL regions were overlapping the 15 most significant DMRs (Table 3). Seven out of 8 sites from a FL region within the GDF7 gene were included in the DMR.F2 (β_{FL}=−0.0029). All sites from a FL region associated with the *PRSS50* gene were part of the DMR.F4 (β_{FL}=−0.0069). Six out of 7 sites from the FL region within the *GPR19* gene were within the DMR.F9 (β_{FL}=0.0076). None of the 36 other FL region were overlapping any folate-related DMRs.

For alcohol consumption, we identified 133 FL regions, 71 regions presenting a positive association and 62 an inverse association. 21 regions were included in alcohol-related DMRs. Among them, 9 were overlapping 6 of the 15 most significant DMRs (Table 4). The situation where two close FL regions were part of the same DMR was observed 3 times in the 15 most significant alcohol-related DMRs. In particular, four and three sites from two FL regions located in chromosome 22 were included in DMR.A11, associated with genes *SMC1B* and *RIBC2*. All the 9 sites from a FL region were included in DMR.A9 (β_{FL}=−0.474).

Graphical representations of the DMRs, the FL regions and their overlap are illustrated for each chromosome in Additional file 6, Figure S4 for dietary folate and Additional file 7, Figure S5 for alcohol intake. For dietary folate, most of FL regions were located in chromosome 3, 22 and chromosome X. A maximum of four DMRs located in the same chromosome was observed for chromosome 2 and 3. As for alcohol intake, DMR and FL showed overlap mostly in chromosomes 6 and 22, with, respectively, 4 and 3 DMRs overlapping FL regions.

Discussion

In this study of women from a large prospective cohort, we investigated the association of dietary folate and alcohol intake with leukocyte DNA methylation via three different approaches. The site-specific analysis aimed at identifying single CpG sites independently

349 from each other, whereas DMRs and FL analyses aimed at identifying regions of CpG sites
350 using the inter-correlation between methylation levels in close sites, thus exploiting the
351 potential of specific regions of the epigenome to show methylation activity related to lifestyle
352 factors.

353 While site-specific analysis showed a lack of association between individual CpG sites and
354 dietary folate, alcohol intake was positively associated with the site cg03199996 and inversely
355 associated with cg07382687. These two sites are located within the body region of the genes
356 *FAMB65C* and *CREB3L2*. The *FAMB65C* gene, also named '*RIPOR3*', is a non-annotated
357 gene. The *CREB3L2* gene encodes a transcriptional activator protein and plays a critical role
358 in cartilage development by activating the transcription of *SEC23A* (Hino et al., 2014).
359 Translocation of *CREB3L2* gene, located on chromosome 7, and the *FUS* gene (fused in
360 sarcoma) located on the chromosome 16 has been found in some tumors, including skin
361 cancer and soft tissue sarcoma (Panagopoulos et al., 2004; Patel et al., 2011).

362 Alcohol is known to alter DNA methylation, mostly because it contributes to deregulation of
363 folate absorption, which can lead to an dysfunction of OCM (Kruman & Fowler, 2014). In our
364 study, alcohol intake was associated with 90 DMRs, some of which may have a role in
365 specific carcinogenesis processes. For example, alcohol intake was inversely associated with
366 methylation levels in DMR.A64 related to the *MLH1* gene, which is frequently mutated in
367 hereditary nonpolyposis colon cancer (HNPCC) (Peltomaki & de la Chapelle, 1997). A
368 positive association between alcohol intake and methylation in the DMR .A79 was related to
369 the *TSPAN32* (tetraspanin 32) gene, also known as the *TSSC6* gene, which is one of the
370 several tumor suppressor genes located at locus 11p15.5 in the imprinted gene domain of
371 chromosome 11 (Lee et al., 1999). This locus has been associated with adrenocortical
372 carcinoma, lung, ovarian and breast cancers. Methylations within DMR.A1 was positively
373 associated with alcohol intake, and the related *GSDMD* gene has also been suggested to act as
374 a tumor suppressor (Saeki et al., 2009). Alcohol intake was also positively associated with
375 DMR.A6 related to the gene *ADAM32*, which encodes a protein involved in diverse biological
376 processes, such as brain development, fertilization, tumor development and inflammation
377 (O'Leary et al., 2016).

378 Several genes, associated with the 24 DMRs identified in our study for dietary folate, were
379 possibly involved in biological processes leading to carcinogenesis. For example, dietary
380 folate was positively associated with methylation in DMR.F16 related to the *RTKN* (rhotekin)
381 gene, which interacts with GTP-bound Rho proteins. Rho proteins regulate many important

382 cellular processes, including cell growth and transformation, cytokinesis, transcription, and
383 smooth muscle contraction. Dysregulation of the Rho signal transduction pathway has been
384 implicated in many forms of cancer such as bladder cancer, gastric cancer and breast cancer
385 (M. Chen, Bresnick, & O'Connor, 2012; Fan et al., 2005). Dietary folate was also associated
386 with methylation levels in DMR.F1 and DMR.F5 within the *HOXA4*, *HOXA5* and *HOXA6*
387 genes, members of the HOX family, known to be associated with cellular differentiation
388 (Seifert, Werheid, Knapp, & Tobiasch, 2015). Perturbed HOX gene expression has been
389 implicated in multiple cancer types (Shah & Sukumar, 2010). In addition, *HOXA5* may also
390 regulate gene expression and morphogenesis. Methylation of this gene may result in the loss
391 of its expression and, since the encoded protein upregulates the tumor suppressor p53, may
392 play an important role in tumorigenesis (Teo et al., 2016).

393 Results from site-specific and DMR analyses were generated with different analytical
394 strategies: methylation levels in different sites were assumed independent in the former, with
395 linear regression models fitted separately in each CpG site, while in the latter the physical
396 proximity of CpGs was exploited to identify specific regions of the epigenome with similar
397 methylation activity, under the assumption that neighboring CpG sites may share relevant
398 epigenetics information. FL analysis revealed some overlaps with DMRs, particularly for
399 alcohol intake, where 9 FL regions were observed within the 15 most significant DMRs. Yet,
400 DMRs and FL analyses have differences, and their results deserve cautious interpretations.
401 Unlike DMRs, FL does not take into account the physical distance between consecutive sites,
402 and rather uses the order of CpG sites on a given chromosome. Methylation levels within a
403 chromosome were mutually adjusted in FL regression, while in DMR analysis t-test statistics
404 were based on independent associations of methylation levels with folate and alcohol.

405 The association between folate and DNA methylation has been investigated at different stage
406 of human life, in particular during fetal development and elderly, where folate is especially
407 needed. A meta-analysis of mother-offspring pairs estimated the association between maternal
408 plasma folate during pregnancy and DNA methylation in cord blood (Joubert et al., 2016).
409 After FDR correction, maternal plasma folate was positively associated with methylation
410 level at 27 CpG sites and inversely associated with methylation level at 416 CpG sites. None
411 of these sites was observed in any of the 24 DMRs related to dietary folate in the present
412 study. This might be explained by the lack of power to identify specific-sites due to the
413 sample size: over 2,000 samples were included in Joubert' meta-analysis against 450 in our

414 study. Then, different methods were used to assess folate intake, i.e. plasma folate against
415 dietary folate.

416 An intervention study was conducted to evaluate the effects of long-term supplementation
417 with folic acid and vitamin B12 on white blood cell DNA methylation in elderly subjects
418 (Kok et al., 2015). After the intervention of two years, 162 sites were significantly
419 differentially methylated compared to baseline, versus 6 sites only for the placebo group.
420 Folate and vitamin B12 were not significantly associated with methylation level in any CpG
421 sites. Within the same study, 173 and 425 DMRs were identified for folate and vitamin B12,
422 respectively. The gene *HOX4*, which was inversely associated with dietary folate in our study
423 in DMR #5, was the only region overlapping with the first 10 DMRs found in the intervention
424 study (Kok et al., 2015). However, a higher level of folic acid was observed in the
425 intervention study: averages blood folate of 52 and 23 nmol/L in the intervention and placebo
426 groups, respectively, compared to an average blood folate of 15 nmol/L in our study which
427 might partly explained the different findings.

428 Within a recent meta-analysis including 9643 participants of European ancestry, aged 42 to 76
429 years with 54% women (Liu et al., 2016), 363 CpGs sites were significantly associated with
430 alcohol consumption, with 87% of these sites showing inverse associations. In our study, site
431 cg02711608 was part of the 363 identified sites, and was also included in DMR #25
432 associated with gene *SLC1A5*. *SLC1A5* gene encodes a protein which is a sodium-dependent
433 amino acids transporter (Pochini, Scalise, Galluccio, & Indiveri, 2014). The important
434 difference in the number of significant sites between the meta-analysis and the present study
435 might mostly be explained by the larger study population size and the larger levels of alcohol
436 intake observed in the meta-analysis (Liu et al., 2016). Indeed, in the meta-analysis,
437 composed of 46% of men, the medians of alcohol intake ranged from 0 to 14 g/day in the 10
438 European cohorts; while with a median of 3.5 g/day, alcohol intake was quite low in our
439 study, which included only women. Lastly, cohort-specific approaches were used in the meta-
440 analysis to remove technical variability, while the SVA approach was used in our study,
441 which was shown to produce conservative findings compared to other normalizing techniques
442 (Perrier et al., 2018).

443 A major strength of this study was the use of a population of European women from United
444 Kingdom, Germany, Italy, Greece, Netherland and Spain, implying diversity of diet and
445 lifestyle habits. Three approaches were used to evaluate the relationship between dietary
446 folate, alcohol intake and DNA methylation. The comparison between DMR and FL analyses

was particularly relevant to identify regions of the genome associated with dietary folate and alcohol intake. Among the DMRs identified in this study for dietary folate or alcohol intake, several regions were associated with genes potentially implicated in cancer development, such as *RTKN,* the *HOX* family of genes and the two tumor suppressor genes *GSDMD* and *TSPAN32*. Our findings need confirmation in other populations. The low number of significant CpGs sites identified in this site-specific analysis may mostly be explained by the relatively low sample size (n=450).

Conclusion

Alcohol intake was associated with methylation levels at two CpG sites. Evidence from DMRs and FL analysis indicated that both dietary folate and alcohol intakes might be associated with alteration of DNA methylation levels in localized regions. Folate and alcohol are known to be associated with breast cancer but also to have a mutually antagonistic role in the one-carbon metabolism. In some regions identified by DMRs or FL analysis, mapped genes are known to act as tumor suppressor such as the *GSDMD* and *HOXA5* genes. These results were in line with the hypothesis that folate- and alcohol-deregulated epigenetic mechanisms might have a role in the pathogenesis of cancer.

Abbreviations

BC: Breast cancer;

BMI: Body mass index;

DMR: Differentially methylated region;

EPIC: European Prospective Investigation into Cancer and nutrition;

FDR: False discovery rate;

FL: Fused lasso;

HM450K: Illumina Infinium HumanMethylation450K;

MAF: minor allele frequency;

NK: natural killer;

OCM: One-carbon metabolite;

SAM: S-adenosylmethyonine;

475 SVA: Surrogate variables analysis.

476 **Declarations**

477 ***Acknowledgements***

478 The authors would like to thank the financial support provided by La Fondation de France for
479 a doctoral fellowship. They are also grateful for all the women who participated in the EPIC
480 cohort and without whom this work would not have been possible.

481 ***Funding***

482 This work was supported by a doctoral fellowship from 'Fondation de France' (grant number
483 2015 00060737) to FP and the grants from the Institut National du Cancer (INCa, France,
484 2012-070 to IR and ZH), la Ligue nationale contre le cancer (to Z. Herceg). ZH was supported
485 by the European Commission (EC) Seventh Framework Programme (FP7) Translational
486 Cancer Research (TRANSCAN) Framework, the Fondation Association pour la Recherche
487 contre le Cancer (ARC, France). In addition, this study was supported by postdoctoral
488 fellowship to SA from the International Agency for Research on Cancer, partially supported
489 by the EC FP7 Marie Curie Actions – People – Co-funding of regional, national and
490 international programmes (COFUND). SA's work is supported by Cancer Research UK
491 (grant number: C18281/A19169). SA work in the Medical Research Council Integrative
492 Epidemiology Unit at the University of Bristol which is supported by the Medical Research
493 Council and the University of Bristol (grant number: MC_UU_00011/1, MC_UU_00011/4
494 and MC_UU_00011/5).

495 The coordination of EPIC is financially supported by the European Commission (DG-
496 SANCO) and the International Agency for Research on Cancer. The national cohorts are
497 supported by German Cancer Aid, German Cancer Research Center (DKFZ), Federal
498 Ministry of Education and Research (BMBF), Deutsche Krebshilfe, Deutsches
499 Krebsforschungszentrum and Federal Ministry of Education and Research (Germany); the
500 Hellenic Health Foundation (Greece); Associazione Italiana per la Ricerca sul Cancro-AIRC-
501 Italy and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and
502 Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention
503 Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF),
504 Statistics Netherlands (The Netherlands); Health Research Fund (FIS), PI13/00061 to
505 Granada, PI13/01162 to EPIC-Murcia, PI13/02633 to EPIC-Navarra), Regional Governments

of Andalucía, Asturias, Basque Country, Murcia and Navarra, ISCIII RETIC (RD06/0020) (Spain); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 and C8221/A19170 to EPIC-Oxford), Medical Research Council (1000143 to EPIC-Norfolk, MR/M012190/1 to EPIC-Oxford) (UK).

The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Availability of data and materials

For information on how to submit an application for gaining access to EPIC data and/or biospecimens, please follow the instructions at http://epic.iarc.fr/access/index.php

Authors' contributions

FP performed the statistical data analysis and drafted the manuscript. IR and PF developed the concept of the study with FP, and contributed to draft the manuscript. SA and CC were responsible for the technical aspects of DNA methylation acquisition. IR and ZH conceived the epigenetics study in the nested case-control study on breast cancer, and critically reviewed the manuscript. SA, AG and HHV contributed to the interpretation of the results. LB, CV, MM and MJG were involved in the data interpretation. All authors contributed to draft the final versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study was approved by the Ethical Review Board of the International Agency for Research on Cancer, and by the local Ethics Committees in the participating centres. This study was also conducted in accordance with the IARC Ethic Committee (Project No 10-22).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

Ambatipudi, S., Cuenin, C., Hernandez-Vargas, H., Ghantous, A., Le Calvez-Kelm, F., Kaaks, R., . . . Herceg, Z. (2016). Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics, 8*(5), 599-618. doi:10.2217/epi-2016-0001

Ambatipudi, S., Horvath, S., Perrier, F., Cuenin, C., Hernandez-Vargas, H., Le Calvez-Kelm, F., . . . Herceg, Z. (2017). DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *European Journal of Cancer, 75*, 299-307. doi:http://dx.doi.org/10.1016/j.ejca.2017.01.014

Ba, Y., Yu, H., Liu, F., Geng, X., Zhu, C., Zhu, Q., . . . Zhang, Y. (2011). Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II in maternal and cord blood. *Eur J Clin Nutr, 65*(4), 480-485. doi:10.1038/ejcn.2010.294

Baglietto, L., English, D. R., Gertig, D. M., Hopper, J. L., & Giles, G. G. (2005). Does dietary folate intake modify effect of alcohol consumption on breast cancer risk? Prospective cohort study. *Bmj, 331*(7520), 807. doi:10.1136/bmj.38551.446470.06

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., & Le, J. M. (2011). High density DNA methylation array with single CpG site resolution. *Genomics, 98*. doi:10.1016/j.ygeno.2011.07.007

Bolstad, B. M. (2001). *Probe level quantile normalization of high density oligonucleotide array data*. Retrieved from http://bmbolstad.com/stuff/qnorm.pdf

Bouckaert, K. P., Slimani, N., Nicolas, G., Vignat, J., Wright, A. J., Roe, M., . . . Finglas, P. M. (2011). Critical evaluation of folate data in European and international databases: recommendations for standardization in international nutritional studies. *Mol Nutr Food Res, 55*(1), 166-180. doi:10.1002/mnfr.201000391

Candès, E. J., Wakin, M. B., & Boyd, S. P. (2008). Enhancing Sparsity by Reweighted ℓ 1 Minimization. *Journal of Fourier Analysis and Applications, 14*(5), 877-905. doi:10.1007/s00041-008-9045-x

Chen, M., Bresnick, A. R., & O'Connor, K. L. (2012). Coupling S100A4 to Rhotekin alters Rho signaling output in breast cancer cells. *Oncogene, 32*, 3754. doi:10.1038/onc.2012.383

https://www.nature.com/articles/onc2012383#supplementary-information

Chen, Y. A., Lemire, M., Choufani, S., Butcher, D. T., Grafodatskaya, D., Zanke, B. W., . . . Weksberg, R. (2013). Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics, 8*(2), 203-209. doi:10.4161/epi.23470

Cole, S. R., Platt, R. W., Schisterman, E. F., Chu, H., Westreich, D., Richardson, D., & Poole, C. (2010). Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology, 39*(2), 417-420. doi:10.1093/ije/dyp334

Du, P., Kibbe, W. A., & Lin, S. M. (2008). lumi: a pipeline for processing Illumina microarray. *Bioinformatics, 24*(13), 1547-1548. doi:10.1093/bioinformatics/btn224

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., & Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics, 11*, 587-587. doi:10.1186/1471-2105-11-587

Fan, J., Ma, L.-J., Xia, S.-J., Yu, L., Fu, Q., Wu, C.-Q., . . . Tang, X.-D. (2005). Association between clinical characteristics and expression abundance of RTKN gene in human bladder carcinoma tissues from Chinese patients. *Journal of Cancer Research and Clinical Oncology, 131*(3), 157-162. doi:10.1007/s00432-004-0638-8

Ferrari, P., Jenab, M., Norat, T., Moskal, A., Slimani, N., Olsen, A., . . . Riboli, E. (2007). Lifetime and baseline alcohol intake and risk of colon and rectal cancers in the European prospective investigation into cancer and nutrition (EPIC). *Int J Cancer, 121*(9), 2065-2072. doi:10.1002/ijc.22966

Hastie, T. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*.

585 Heyn, H., Li, N., Ferreira, H. J., Moran, S., Pisano, D. G., & Gomez, A. (2012). Distinct DNA
586 methylomes of newborns and centenarians. *Proc Natl Acad Sci USA, 109*.
587 doi:10.1073/pnas.1120658109
588 Hino, K., Saito, A., Kido, M., Kanemoto, S., Asada, R., Takai, T., . . . Imaizumi, K. (2014).
589 Master regulator for chondrogenesis, Sox9, regulates transcriptional activation of the
590 endoplasmic reticulum stress transducer BBF2H7/CREB3L2 in chondrocytes. *J Biol*
591 *Chem, 289*(20), 13810-13820. doi:10.1074/jbc.M113.543322
592 Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock
593 theory of ageing. *Nat Rev Genet, 19*(6), 371-384. doi:10.1038/s41576-018-0004-3
594 Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J.,
595 Nelson, H. H., . . . Kelsey, K. T. (2012). DNA methylation arrays as surrogate
596 measures of cell mixture distribution. *BMC Bioinformatics, 13*(1), 1-16.
597 doi:10.1186/1471-2105-13-86
598 Huber, W., Carey, V. J., Gentleman, R., Anders, S., Carlson, M., Carvalho, B. S., . . .
599 Morgan, M. (2015). Orchestrating high-throughput genomic analysis with
600 Bioconductor. *Nat Methods, 12*(2), 115-121. doi:10.1038/nmeth.3252
601 Illumina (Producer). (2011). GenomeStudio/BeadStudio Software Methylation Module.
602 Joehanes, R., Just, A. C., Marioni, R. E., Pilling, L. C., Reynolds, L. M., Mandaviya, P. R., . . .
603 London, S. J. (2016). Epigenetic Signatures of Cigarette Smoking. *Circ Cardiovasc*
604 *Genet, 9*(5), 436-447. doi:10.1161/circgenetics.116.001506
605 Joubert, B. R., den Dekker, H. T., Felix, J. F., Bohlin, J., Ligthart, S., Beckett, E., . . . London,
606 S. J. (2016). Maternal plasma folate impacts differential DNA methylation in an
607 epigenome-wide meta-analysis of newborns. *Nat Commun, 7*, 10577.
608 doi:10.1038/ncomms10577
609 Kok, D. E., Dhonukshe-Rutten, R. A., Lute, C., Heil, S. G., Uitterlinden, A. G., van der Velde,
610 N., . . . Steegenga, W. T. (2015). The effects of long-term daily folic acid and vitamin
611 B12 supplementation on genome-wide DNA methylation in elderly subjects. *Clin*
612 *Epigenetics, 7*, 121. doi:10.1186/s13148-015-0154-5
613 Kruman, II, & Fowler, A. K. (2014). Impaired one carbon metabolism and DNA methylation in
614 alcohol toxicity. *J Neurochem, 129*(5), 770-780. doi:10.1111/jnc.12677
615 Lee, M. P., Brandenburg, S., Landes, G. M., Adams, M., Miller, G., & Feinberg, A. P. (1999).
616 Two novel genes in the center of the 11p15 imprinted domain escape genomic
617 imprinting. *Hum Mol Genet, 8*(4), 683-690.
618 Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva
619 package for removing batch effects and other unwanted variation in high-throughput
620 experiments. *Bioinformatics, 28*(6), 882-883. doi:10.1093/bioinformatics/bts034
621 Leek, J. T., & Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by
622 surrogate variable analysis. *PLoS Genet, 3*. doi:10.1371/journal.pgen.0030161
623 Leek, J. T., & Storey, J. D. (2008). A general framework for multiple testing dependence.
624 *Proc Natl Acad Sci U S A, 105*(48), 18718-18723. doi:10.1073/pnas.0808709105
625 Liu, C., Marioni, R. E., Hedman, A. K., Pfeiffer, L., Tsai, P. C., Reynolds, L. M., . . . Levy, D.
626 (2016). A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry*.
627 doi:10.1038/mp.2016.192
628 Mason, J. B., & Choi, S.-W. (2005). Effects of alcohol on folate metabolism: implications for
629 carcinogenesis. *Alcohol, 35*(3), 235-241.
630 doi:https://doi.org/10.1016/j.alcohol.2005.03.012
631 Matejcic, M., de Batlle, J., Ricci, C., Biessy, C., Perrier, F., Huybrechts, I., . . . Chajes, V.
632 (2017). Biomarkers of folate and vitamin B12 and breast cancer risk: report from the
633 EPIC cohort. *Int J Cancer, 140*(6), 1246-1259. doi:10.1002/ijc.30536
634 Niculescu, M. D., & Haggarty, P. (2011). *Nutrition in Epigenetics*: Wiley.
635 O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K.
636 D. (2016). Reference sequence (RefSeq) database at NCBI: current status,
637 taxonomic expansion, and functional annotation. *Nucleic Acids Res, 44*(D1), D733-
638 745. doi:10.1093/nar/gkv1189

Panagopoulos, I., Storlazzi, C. T., Fletcher, C. D., Fletcher, J. A., Nascimento, A., Domanski, H. A., . . . Mertens, F. (2004). The chimeric FUS/CREB3l2 gene is specific for low-grade fibromyxoid sarcoma. *Genes Chromosomes Cancer, 40*(3), 218-228. doi:10.1002/gcc.20037

Patel, R. M., Downs-Kelly, E., Dandekar, M. N., Fanburg-Smith, J. C., Billings, S. D., Tubbs, R. R., & Goldblum, J. R. (2011). FUS (16p11) gene rearrangement as detected by fluorescence in-situ hybridization in cutaneous low-grade fibromyxoid sarcoma: a potential diagnostic tool. *Am J Dermatopathol, 33*(2), 140-143. doi:10.1097/IAE.0b013e318176de80

Peltomaki, P., & de la Chapelle, A. (1997). Mutations predisposing to hereditary nonpolyposis colorectal cancer. *Adv Cancer Res, 71*, 93-119.

Perrier, F., Novoloaca, A., Ambatipudi, S., Baglietto, L., Ghantous, A., Perduca, V., . . . Ferrari, P. (2018). Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenetics, 10*(1), 38. doi:10.1186/s13148-018-0471-6

Peters, T. J., Buckley, M. J., Statham, A. L., Pidsley, R., Samaras, K., V Lord, R., . . . Molloy, P. L. (2015). De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin, 8*(1), 1-16. doi:10.1186/1756-8935-8-6

Pochini, L., Scalise, M., Galluccio, M., & Indiveri, C. (2014). Membrane transporters for the special amino acid glutamine: structure/function relationships and relevance to human health. *Frontiers in Chemistry, 2*(61). doi:10.3389/fchem.2014.00061

Riboli, E., Hunt, K. J., Slimani, N., Ferrari, P., Norat, T., Fahey, M., . . . Saracci, R. (2002). European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr, 5*(6B), 1113-1124. doi:10.1079/phn2002394

Saeki, N., Usui, T., Aoyagi, K., Kim, D. H., Sato, M., Mabuchi, T., . . . Sasaki, H. (2009). Distinctive expression and function of four GSDM family genes (GSDMA-D) in normal and malignant upper gastrointestinal epithelium. *Genes Chromosomes Cancer, 48*(3), 261-271. doi:10.1002/gcc.20636

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2*. doi:10.2307/3002019

Seifert, A., Werheid, D. F., Knapp, S. M., & Tobiasch, E. (2015). Role of Hox genes in stem cell differentiation. *World J Stem Cells, 7*(3), 583-595. doi:10.4252/wjsc.v7.i3.583

Shah, N., & Sukumar, S. (2010). The Hox genes and their roles in oncogenesis. *Nat Rev Cancer, 10*(5), 361-371. doi:10.1038/nrc2826

Slimani, N., Deharveng, G., Unwin, I., Southgate, D. A., Vignat, J., Skeie, G., . . . Riboli, E. (2007). The EPIC nutrient database project (ENDB): a first attempt to standardize nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr, 61*(9), 1037-1056. doi:10.1038/sj.ejcn.1602679

Slimani, N., Ferrari, P., Ocke, M., Welch, A., Boeing, H., Liere, M., . . . Riboli, E. (2000). Standardization of the 24-hour diet recall calibration method used in the european prospective investigation into cancer and nutrition (EPIC): general concepts and preliminary results. *Eur J Clin Nutr, 54*(12), 900-917.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet and Mol Biol, 3*.

Smyth, G. K. (2005). Limma: linear models for microarray data. In R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, & W. Huber (Eds.), *Bioinformatics and Computational Biology Solutions Using R. and Bioconductor*. New York: Springer.

Szyf, M. (2011). The implications of DNA methylation for toxicology: toward toxicomethylomics, the toxicology of DNA methylation. *Toxicol Sci, 120*(2), 235-255. doi:10.1093/toxsci/kfr024

Teegarden, D., Romieu, I., & Lelievre, S. A. (2012). Redefining the impact of nutrition on breast cancer incidence: is epigenetics involved? *Nutr Res Rev, 25*(1), 68-95. doi:10.1017/s0954422411000199

693 Teo, W. W., Merino, V. F., Cho, S., Korangath, P., Liang, X., Wu, R.-c., . . . Sukumar, S.
694 (2016). HOXA5 determines cell fate transition and impedes tumor initiation and
695 progression in breast cancer through regulation of E-cadherin and CD24. *Oncogene,*
696 *35*(42), 5539-5551. doi:10.1038/onc.2016.95
697 Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., &
698 Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe
699 design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics, 29*(2),
700 189-196. doi:10.1093/bioinformatics/bts680
701 Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and
702 smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*
703 *(Statistical Methodology), 67*(1), 91-108. doi:doi:10.1111/j.1467-9868.2005.00490.x
704 Wareham, N. J., Jakes, R. W., Rennie, K. L., Schuit, J., Mitchell, J., Hennings, S., & Day, N.
705 E. (2003). Validity and repeatability of a simple index derived from the short physical
706 activity questionnaire used in the European Prospective Investigation into Cancer and
707 Nutrition (EPIC) study. *Public Health Nutr, 6*(4), 407-413. doi:10.1079/phn2002439
708 Zhang, S., Hunter, D. J., Hankinson, S. E., Giovannucci, E. L., Rosner, B. A., Colditz, G. A., .
709 . . Willett, W. C. (1999). A prospective study of folate intake and the risk of breast
710 cancer. *Jama, 281*(17), 1632-1637.
711 Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American*
712 *Statistical Association, 101*(476), 1418-1429. doi:10.1198/016214506000000735
713

714

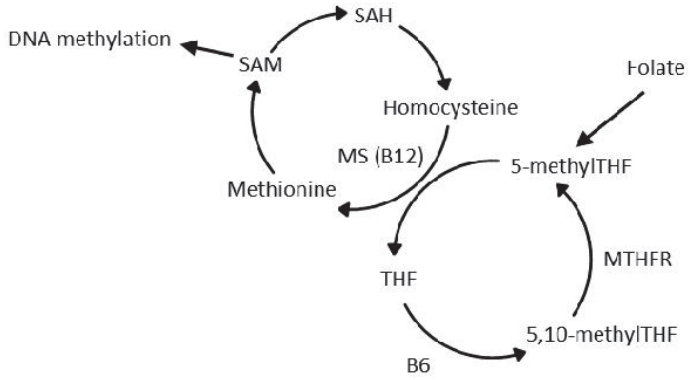
715 **Table 1.** Characteristics of the study population (n=450).

	Mean (SD[1])	Min-Max
Age at blood collection (year)	52 (9)	26-73
Weight (kg)	66 (11)	40-103
Height (cm)	161 (7)	143-196
BMI (kg/m^2)	26 (4)	16-43
Alcohol intake (g/day)	8 (12)	0-72
Blood folate level (nmol/L)	15 (10)	1-89
Dietary folate (µg/day)	270 (106)	91-1012
Cd8t (%)	7.5 (4)	0-23
Cd4t (%)	13.5 (5)	0-34
Natural killer (%)	6.7 (5)	0-27
B cells (%)	6.1 (2)	0-17
Monocytes (%)	5.7 (3)	0-17
Granulocytes (%)	60.8 (9)	27-85
	N	%
Menopausal status:		
- Pre-menopause	186	41.3
- Post-menopause	264	58.7
Smoking status		
- Never	250	55.6
- Former	93	20.7
- Smoker	104	23.1
- Missing	3	0.7
Physical activity Index[2]		
- Inactive	99	22.0
- Moderate inactive	187	41.5
- Moderate active	75	16.7
- Active	78	10.7
- Missing	11	2.4

716 [1] SD=standard deviation, reported for continuous variables only;
717 [2] (Wareham et al., 2003).

718

719 **Figure 1.** Diagram of the one-carbon metabolism pathway.



720

721 MS: methionine synthase; MTHFR: methylenetetrahydrofolate reductase; THF:
722 tetrahydrofolate; SAH: S-adenosylhomocysteine; SAM: S-adenosylmethonine.
723

724 **Table 2.** CpG site-specific models results for the significant CpG sites for alcohol intake[1].

725

	CpGs names	Alcohol intake $\beta_{(1SD)}$[2]	Alcohol intake q_{val}[3]	CpGs characteristics Associated genes	CpGs characteristics Gene region[4]	CpGs characteristics Island[5]	CpGs characteristics Chr
1	cg07382687	-0,0389	0.049	CREB3L2	Body	Open Sea	7
2	cg03199996	0,0442	0.049	FAM65C	Body	Open Sea	20

726 [1] Adjusted for recruitment centre, age at recruitment, menopausal status and level of different lymphocyte
727 subtypes;
728 [2] Coefficients for 1 standard deviation alcohol intake (SD=11.8);
729 [3] False discovery rate (FDR) adjusted p-values;
730 [4] Gene region feature category describing the CpG position, from UCSC. TSS200: 200 bases upstream of the
731 transcriptional start site (TSS); TSS1500: 1500 bases upstream of the TSS; 5'UTR: Within the 5' untranslated
732 region, between the TSS and the ATG start site; Body: Between the ATG and stop codon; irrespective of the
733 presence of introns, exons, TSS, or promoters; 3'UTR: Between the stop codon and poly A signal;
734 [5] The location of the CpG relative to the CpG island. Shore: 0–2 kb from island; Shelf: 2–4 kb from island; N:
735 upstream (5') of CpG island; S: downstream (3') of CpG island; Open Sea: Isolated CpGs in the genome.

Table 3. The 15 most significant DMRs associated with dietary folate out of 24 significant DMRs [1].

	DMRs characteristics				CpGs characteristics		Fused Lasso	
	Associated genes	Gene regions	hg19coord	Sites[2]	q_{DMR}[3]	β_{DMR}[4]	overlap[5]	β_{FL}[6]
F1	HOXA5,HOXA6	1stExon, 5'UTR, TSS200, TSS1500, 3'UTR, Body	chr7:27183133-27185512	49	1,3E-13	0,054		
F2	GDF7	Body	chr2:20869434-20871401	8	1,4E-08	-0,094	7/8	-0.0029
F3	CYP1A1	TSS1500	chr15:75018731-75019376	13	2,4E-05	0,041		
F4	PRSS50	Body, 1stExon, 5'UTR, TSS200, TSS1500	chr3:46759096-46759698	9	2,4E-04	-0,056	4/4	-0.0069
F5	HOXA4	1stExon, 5'UTR, TSS200, TSS1500	chr7:27170241-27171154	14	5,8E-04	-0,047		
F6	SYNGAP1	Body	chr6:33401192-33401542	6	1,0E-03	0,024		
F7	ZNF833	TSS1500, TSS200, Body	chr19:11784514-11785337	13	1,1E-03	-0,036		
F8	LAMB2	1stExon, 5'UTR, TSS200, TSS1500	chr3:49170496-49170849	6	3,1E-03	-0,035		
F9	GPR19	5'UTR, 1stExon, TSS200, TSS1500	chr12:12848977-12849588	9	3,7E-03	0,065	6/7	0.0076
F10	MTMR15	TSS1500, TSS200, 5'UTR, 1stExon	chr15:31195612-31196075	7	4,0E-03	-0,050		
F11	KCNE1	5'UTR, 1stExon, TSS200, TSS1500	chr21:35831871-35832364	8	4,2E-03	0,054		
F12	TNXB	Body	chr6:32054659-32055474	20	7,2E-03	-0,036		
F13	TERT	Body	chr5:1269992-1270152	3	7,2E-03	0,033		
F14	C2orf27A	5'UTR	chr2:132481613-132481826	2	1,7E-02	0,089		
F15	ANKRD44	Body	chr2:198029141-198029332	3	2,1E-02	-0,052		

[1] Adjusted for recruitment centre, age at recruitment, menopausal status and level of different lymphocyte subtypes;
[2] Number of sites located in DMRs significant for dietary folate;
[3] Minimum dietary folate q-values of sites located in the DMRs (FDR correction);
[4] Absolute maximum of dietary folate coefficient of sites located in the DMRs;
[5] Number of sites from the FL region overlapping the DMR / number of sites in the FL region;
[6] Dietary folate coefficient of sites located in the FL region.

25

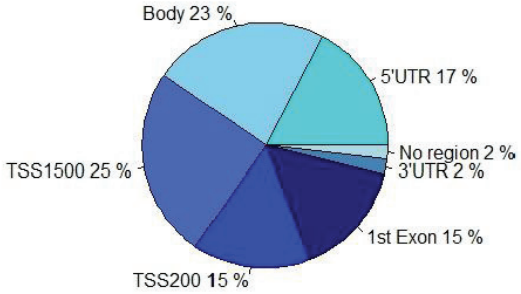
Table 4. The 15 most significant DMRs associated with alcohol out of 90 significant DMRs[1].

	DMRs characteristics				CpGs characteristics		Fused Lasso	
	Associated genes	Gene regions	hg19coord	Sites[2]	q_{DMR}[3]	β_{DMR}[4]	overlap[5]	β_{FL}[6]
A1	GSDMD	TSS1500, TSS200, 5'UTR, 1stExon	chr8:144635260-144636462	9	4.7E-14	0.0017		
A2			chr6:31650735-31651362	21	1.8E-13	0.0015	2/2, 2/2	0.390
A3	TRIM4	Body, 1stExon, 5'UTR, TSS200, TSS1500	chr7:99516603-99517509	14	3.0E-06	0.0015		
A4	RGL3	Body	chr19:11517079-11517436	5	3.3E-06	0.0017		
A5	COL9A3	TSS1500	chr20:61446962-61447992	32	4.8E-06	-0.0010	4/4	-1.027
A6	ADAM32	TSS1500, TSS200, 1stExon, 5'UTR, Body	chr8:38964500-38965492	10	1.3E-04	0.0012		
A7	C21orf56	5'UTR, 1stExon, TSS1500	chr21:47604052-47605174	8	1.5E-04	0.0027		
A8			chr2:118616155-118616576	5	1.9E-04	0.0016	5/7	0.514
A9	LTB4R2, LTB4R, CIDEB	Body, 1stExon, TSS1500, 5'UTR, TSS200	chr14:24780404-24780926	10	2.3E-04	-0.0010	9/9	-0.474
A10	PTDSS2	Body	chr11:457256-457304	3	3.0E-04	0.0009		
A11	SMC1B, RIBC2	Body, TSS1500, 1stExon, TSS200, 5'UTR	chr22:45808669-45810043	16	3.0E-04	0.0016	4/4, 3/3	0.332
A12			chr10:72013286-72013397	2	8.4E-04	-0.0012		
A13	TRAF3	Body	chr14:103366987-103367858	5	1.4E-03	0.0011		
A14	C22orf27	TSS1500, TSS200, Body	chr22:31317764-31318546	12	1.4E-03	0.0013	4/4, 2/2	0.641
A15	S100A13, S100A1	5'UTR, 1stExon, TSS1500, TSS200	chr1:153599479-153600156	8	3.0E-03	0.0016		

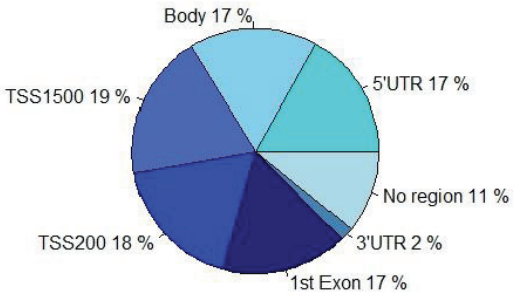
[1] Adjusted for recruitment centre, age at recruitment, menopausal status and level of different lymphocyte subtypes;
[2] Number of sites located in DMRs significant for alcohol;
[3] Minimum alcohol q-values of sites located in the DMRs (FDR correction);
[4] Absolute maximum of alcohol coefficient of sites located in the DMRs;
[5] Number of sites from the FL region overlapping the DMR / number of sites in the FL region, appears twice if two FL regions are included in the DMR;
[6] Alcohol coefficient of sites located in the FL region or average of alcohol coefficients if two FL regions are included in a DMR.

26

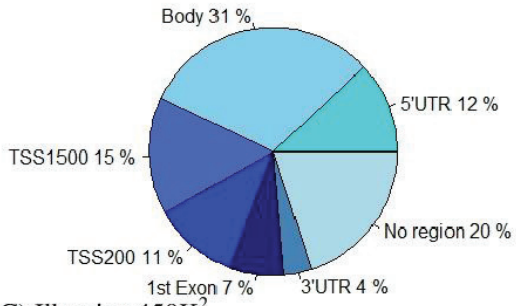
Figure 2. Repartition of gene regions[1] among DMRs compare to their repartition within the Illumina 450K[2].



A) DMRs significant for folate



B) DMRs significant for alcohol

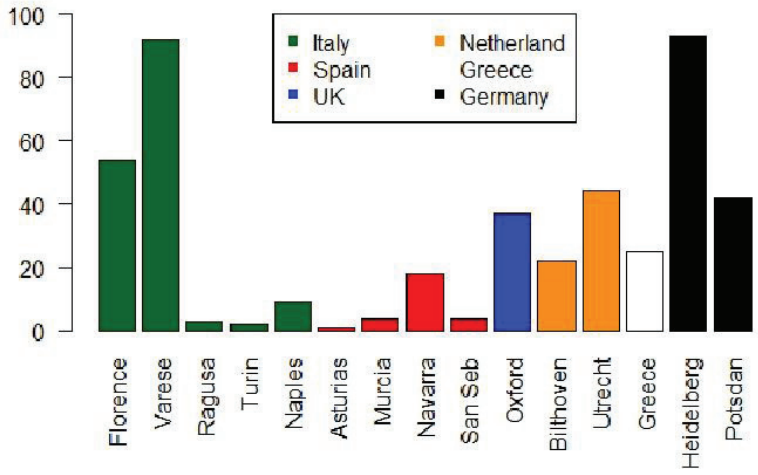


C) Illumina 450K[2]

[1] Gene region feature category describing the CpG position, from UCSC. TSS200: 200 bases upstream of the transcriptional start site (TSS); TSS1500: 1500 bases upstream of the TSS; 5'UTR: Within the 5' untranslated region, between the TSS and the ATG start site; Body: Between the ATG and stop codon; irrespective of the presence of introns, exons, TSS, or promoters; 3'UTR: Between the stop codon and poly A signal.
[2] The repartition of CpGs sites was done among the 421,583 sites included in this study.

Supplementary Materials for "Association of leukocyte DNA methylation changes with dietary folate and alcohol intake in the EPIC Study"

Additional file 1. Sample size by recruitment centers.



Additional file 2. DMRs associated with dietary folate (log)[1].

	Associated genes	Gene regions	hg19coord	Sites[2]	q_{DMR}[3]	β_{DMR}[4]
1	HOXA5,HOXA6	1stExon,5'UTR,TSS200,TSS1500,3'UTR,Body	chr7:27183133-27185512	49	1.3E-13	0.054
2	GDF7	Body	chr2:20869434-20871401	8	1.4E-08	-0.094
3	CYP1A1	TSS1500	chr15:75018731-75019376	13	2.4E-05	0.041
4	PRSS50	Body,1stExon,5'UTR,TSS200,TSS1500	chr3:46759096-46759698	9	2.4E-04	-0.056
5	HOXA4	1stExon,5'UTR,TSS200,TSS1500	chr7:27170241-27171154	14	5.8E-04	-0.047
6	SYNGAP1	Body	chr6:33401192-33401542	6	1.0E-03	0.024
7	ZNF833	TSS1500,TSS200,Body	chr19:11784514-11785337	13	1.1E-03	-0.036
8	LAMB2	1stExon,5'UTR,TSS200,TSS1500	chr3:49170496-49170849	6	3.1E-03	-0.035
9	GPR19	5'UTR,1stExon,TSS200,TSS1500	chr12:12848977-12849588	9	3.7E-03	0.065
10	MTMR15	TSS1500,TSS200,5'UTR,1stExon	chr15:31195612-31196075	7	4.0E-03	-0.050
11	KCNE1	5'UTR,1stExon,TSS200,TSS1500	chr21:35831871-35832364	8	4.2E-03	0.054
12	TNXB	Body	chr6:32054659-32055474	20	7.2E-03	-0.036
13	TERT	Body	chr5:1269992-1270152	3	7.2E-03	0.033
14	C2orf27A	5'UTR	chr2:132481613-132481826	2	1.7E-02	0.089
15	ANKRD44	Body	chr2:198029141-198029332	3	2.1E-02	-0.052
16	RTKN	Body,TSS1500	chr2:74668072-74668286	2	2.9E-02	0.023
17	PTPRN2	Body	chr7:157406607-157406737	2	3.0E-02	0.033
18	PANX1	Body	chr11:93862716-93862749	2	3.5E-02	0.027
19	CDH3	1stExon,5'UTR	chr16:68678841-68678998	2	3.7E-02	-0.024
20			chr8:713162-713216	3	3.8E-02	0.034
21	LEFTY2	TSS1500	chr1:226129481-226129561	2	3.9E-02	-0.042
22	EML1	TSS1500	chr14:100259329-100259352	3	4.1E-02	-0.037
23	CHRNB1	Body	chr17:7350244-7350282	2	4.1E-02	-0.029
24	PAQR9	TSS1500	chr3:142682652-142682682	2	4.2E-02	-0.025

[1] Adjusted for recruitment centre, age at recruitment, menopausal status, level of different lymphocyte subtypes and BC status;
[2] Number of sites located in DMRs significant for dietary folate;
[3] Minimum dietary folate q-values of sites located in the DMRs (FDR correction);
[4] Absolute maximum of dietary folate coefficient of sites located in the DMRs.

Additional file 3. DMRs associated with alcohol intake[1].

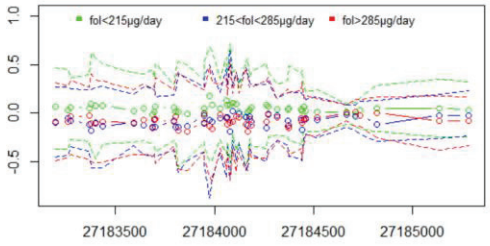
	Associated genes	Gene regions	hg19coord	Sites[2]	q_{DMR}[3]	β_{DMR}[4]
1	GSDMD	TSS1500, TSS200, 5'UTR, 1stExon	chr8:144635260-144636462	9	4,7E-14	0,0017
2			chr6:31650735-31651362	21	1,8E-13	0,0015
3	TRIM4	Body, 1stExon, 5'UTR, TSS200, TSS1500	chr7:99516603-99517509	14	3,0E-06	0,0015
4	RGL3	Body	chr19:11517079-11517436	5	3,3E-06	0,0017
5	COL9A3	TSS1500	chr20:61446962-61447992	32	4,8E-06	-0,0010
6	ADAM32	TSS1500, TSS200, 1stExon, 5'UTR, Body	chr8:38964500-38965492	10	1,3E-04	0,0012
7	C21orf56	5'UTR, 1stExon, TSS1500	chr21:47604052-47605174	8	1,5E-04	0,0027
8			chr2:118616155-118616576	5	1,9E-04	0,0016
9	LTB4R2, LTB4R, CIDEB	Body, 1stExon, TSS1500, 5'UTR, TSS200	chr14:24780404-24780926	10	2,3E-04	-0,0010
10	PTDSS2	Body	chr11:457256-457304	3	3,0E-04	0,0009
11	SMC1B, RIBC2	Body, TSS1500, 1stExon, TSS200, 5'UTR	chr22:45808669-45810043	16	3,0E-04	0,0016
12			chr10:72013286-72013397	2	8,4E-04	-0,0012
13	TRAF3	Body	chr14:103366987-103367858	5	1,4E-03	0,0011
14	C22orf27	TSS1500, TSS200, Body	chr22:31317764-31318546	12	1,4E-03	0,0013
15	S100A13, S100A1	5'UTR, 1stExon, TSS1500, TSS200	chr1:153599479-153600156	8	3,0E-03	0,0016
16	VARS	Body	chr6:31760521-31761076	12	3,1E-03	0,00059
17			chr13:20781097-20781165	3	3,2E-03	0,00096
18			chr6:5783800-5783863	2	3,3E-03	0,00137
19	TSSK6, NDUFA13	1stExon, TSS1500, 5'UTR, TSS200	chr19:19625761-19626599	6	3,8E-03	0,00082
20			chr6:27637302-27637537	4	3,9E-03	0,00122
21	THUMPD3	TSS1500, TSS200, 5'UTR, 1stExon	chr3:9404422-9405070	9	4,5E-03	0,00112
22	NKX2-6	Body, TSS200, TSS1500	chr8:23562918-23564294	12	4,9E-03	0,00091
23	ATP2B2	3'UTR, Body	chr3:10370264-10370704	4	5,5E-03	0,00089
24			chr6:30094947-30095802	26	5,5E-03	-0,00091
25	SLC1A5	Body, 5'UTR, 1stExon, TSS200	chr19:47287778-47288263	6	5,5E-03	-0,00072
26	PCSK4	Body	chr19:1486986-1487605	5	5,7E-03	0,00121
27	IRF6	5'UTR, 1stExon, TSS200, TSS1500	chr1:209979111-209979779	8	5,7E-03	-0,00175
28	C7orf16	TSS1500, TSS200, 5'UTR, 1stExon	chr7:31726494-31726912	6	6,1E-03	-0,00121
29	ADM2, MIOX	3'UTR, TSS1500, TSS200, 5'UTR, 1stExon	chr22:50924745-50925337	5	6,3E-03	0,00127
30	DUPD1	Body	chr10:76803669-76803925	3	6,4E-03	0,00074
31	GOLPH3L	TSS1500	chr1:150670196-150670422	2	6,9E-03	0,00193
32	FLJ44606	5'UTR, TSS200, TSS1500	chr5:126408756-126409573	13	7,4E-03	0,00145
33	C2orf27B	5'UTR, 1stExon, TSS200, TSS1500	chr2:132558939-132559484	6	7,4E-03	-0,00128
34			chr7:157294107-157294502	5	7,6E-03	0,00115

35	ALDH7A1	1stExon, TSS200	chr5:125930870-125931275	6	7,9E-03	0,00101
36	HGFAC	Body	chr4:3449663-3449904	3	8,6E-03	0,00106
37			chr2:129494526-129494877	4	8,7E-03	-0,00061
38	GABBR1	Body	chr6:29599012-29599390	10	8,9E-03	0,00064
39	SLC39A4	Body	chr8:145638202-145639181	5	8,9E-03	0,00131
40			chr11:8361190-8361530	3	8,9E-03	0,00132
41	C17orf98	1stExon, TSS200	chr17:36997449-36997740	5	8,9E-03	0,00130
42			chr11:128557481-128557965	3	9,6E-03	-0,00089
43	ZAP70	5'UTR, Body	chr2:98340425-98340921	4	1,1E-02	0,00146
44	SFRS8	Body	chr12:132270218-132270829	4	1,2E-02	-0,00121
45	LOC654342	TSS200, TSS1500	chr2:91847976-91848218	8	1,2E-02	-0,00080
46	SNORD46, RPS8, SNORD38A	TSS200, Body, TSS1500	chr1:45242073-45242356	3	1,3E-02	-0,00079
47			chr5:80493-80900	2	1,3E-02	0,00205
48			chr22:50165244-50165512	3	1,4E-02	-0,00133
49	C21orf88	TSS1500	chr21:40985387-40985406	2	1,5E-02	0,00097
50	PPT2, PRRT1	TSS1500, TSS200	chr6:32120623-32121261	23	1,5E-02	-0,00102
51			chr6:167559913-167560727	4	1,6E-02	-0,00141
52	ADAM12	1stExon, 5'UTR	chr10:128076910-128076941	3	1,6E-02	0,00087
53			chr13:110521956-110522297	5	1,7E-02	-0,00172
54	GABRG2	TSS1500	chr5:161494015-161494307	3	1,7E-02	0,00146
55	ATHL1	5'UTR, Body	chr11:289774-290292	4	1,7E-02	0,00249
56	SHANK1	Body	chr19:51171061-51171247	2	1,7E-02	0,00123
57	SST	1stExon, 5'UTR, TSS200	chr3:187388128-187388281	4	1,7E-02	0,00084
58	IQSEC1	Body, 5'UTR, 1stExon, TSS200	chr3:13114343-13114803	6	1,8E-02	-0,00109
59			chr15:96911165-96911531	2	1,8E-02	0,00055
60	ANKRD30B	TSS1500, TSS200, 1stExon, 5'UTR	chr18:14747888-14748298	9	1,9E-02	-0,00108
61	OTUD5	5'UTR, 1stExon, TSS200, TSS1500	chrX:48814580-48815125	9	2,0E-02	-0,00048
62	SDR42E1	5'UTR, 1stExon, TSS200	chr16:82044738-82045183	7	2,1E-02	0,00102
63	ISLR2	Body, 3'UTR	chr15:74427234-74427499	3	2,3E-02	0,00084
64	EPM2AIP1, MLH1	1stExon, TSS1500	chr3:37033625-37033903	5	2,4E-02	-0,00114
65			chr21:44253516-44253611	2	2,5E-02	0,00065
66	TNFRSF19	TSS1500, TSS200, 1stExon, 5'UTR	chr13:24144483-24144985	6	2,5E-02	0,00093
67			chr5:180541684-180541897	3	2,6E-02	0,00099
68	CCDC42B	Body	chr12:113592319-113592619	4	3,0E-02	-0,00068
69	SCARF2	Body	chr22:20783497-20783963	3	3,1E-02	0,00125
70	SHISA7	1stExon	chr19:55953951-55953996	2	3,2E-02	0,00134
71	LOC285796	TSS1500	chr6:163746113-163746175	2	3,3E-02	0,00103
72	COX4I2	TSS200, 1stExon, 5'UTR	chr20:30225517-30225851	6	3,3E-02	0,00073
73	CLDN9	1stExon, 5'UTR	chr16:3062597-3062975	4	3,4E-02	-0,00069
74	MAD1L1	Body	chr7:1896154-1896220	2	3,4E-02	0,00103
75	PARP9, DTX3L	5'UTR, TSS1500	chr3:122281881-122281975	3	3,5E-02	-0,00096

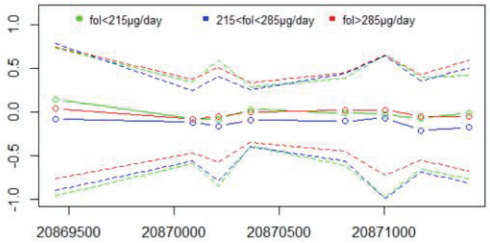
76	PRKCZ	5'UTR, Body	chr1:2058417-2058790	2	3,5E-02	-0,00134
77	HOXC8	TSS1500, TSS200	chr12:54402431-54402717	7	3,6E-02	0,00155
78	ACTC1	Body	chr15:35086890-35086985	4	3,6E-02	0,00106
79	C11orf21, TSPAN32	Body, TSS1500, 1stExon, TSS200	chr11:2322500-2323083	13	3,7E-02	0,00083
80	C6orf123	TSS200, TSS1500	chr6:168197699-168197983	4	3,9E-02	-0,00114
81			chr6:41376604-41376993	3	4,0E-02	-0,00121
82	KRT26	1stExon, 5'UTR	chr17:38928324-38928380	2	4,1E-02	0,00186
83	ASCL2	TSS1500	chr11:2292890-2293117	8	4,3E-02	-0,00089
84	LCE3A	1stExon	chr1:152595322-152595351	2	4,4E-02	-0,00099
85	PIWIL1	TSS200, 1stExon, 5'UTR	chr12:130822603-130822674	3	4,5E-02	-0,00213
86	C1orf163	Body	chr1:53163523-53163758	3	4,5E-02	0,00070
87			chr19:48675305-48676053	2	4,7E-02	0,00051
88	TMEM51	5'UTR	chr1:15541182-15541349	4	4,8E-02	0,00069
89	FKBP11	TSS200, TSS1500	chr12:49319500-49319673	3	4,9E-02	0,00062
90	TRIM69	TSS1500	chr15:45028083-45028098	2	5,0E-02	-0,00032

[1] Adjusted for recruitment centre, age at recruitment, menopausal status, level of different lymphocyte subtypes and BC status;
[2] Number of sites located in DMRs significant for alcohol;
[3] Minimum alcohol q-values of sites located in the DMRs (FDR correction);
[4] Absolute maximum of alcohol coefficient of sites located in the DMRs.

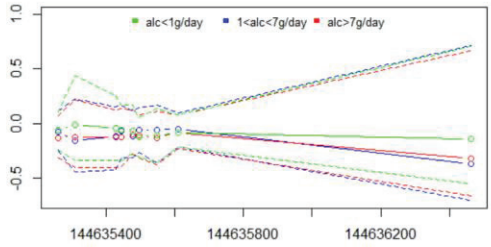
Additional file 4. Graphical representation of the two most significant DMRs of dietary folate and alcohol intake.
The x-axis represents the position (hg 19 coordinates) of the CpGs included in the plotted DMR.
Each tertile of dietary folate, alcohol intake or their interaction are represented by different colors: green for T1, blue for T2 and red for T3. For all the CpGs included in the plotted DMR, the dashed lines are their 1st and 3rd quartiles of methylation levels and the points represent their median values.



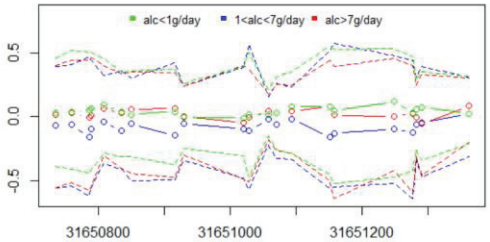
A) Dietary folate: chr7:27183133-27185512



B) Dietary folate: chr2:20869434-20871401

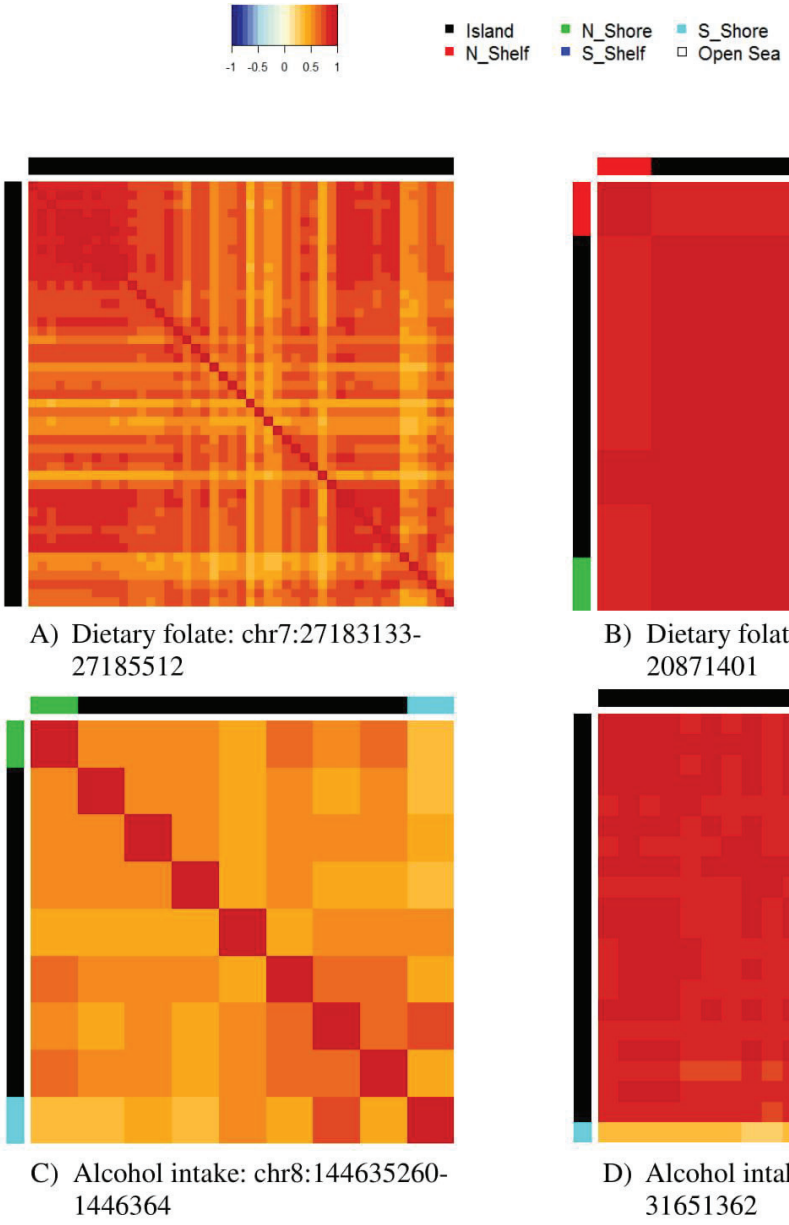


C) Alcohol intake: chr8:144635260-14463646



D) Alcohol intake: chr6:31650735-31651362

Additional file 5. Correlation heatmaps of CpG methylation levels in the two most significant DMR of dietary folate and alcohol intake.



A) Dietary folate: chr7:27183133-27185512

B) Dietary folate: chr2:20869434-20871401

C) Alcohol intake: chr8:144635260-1446364

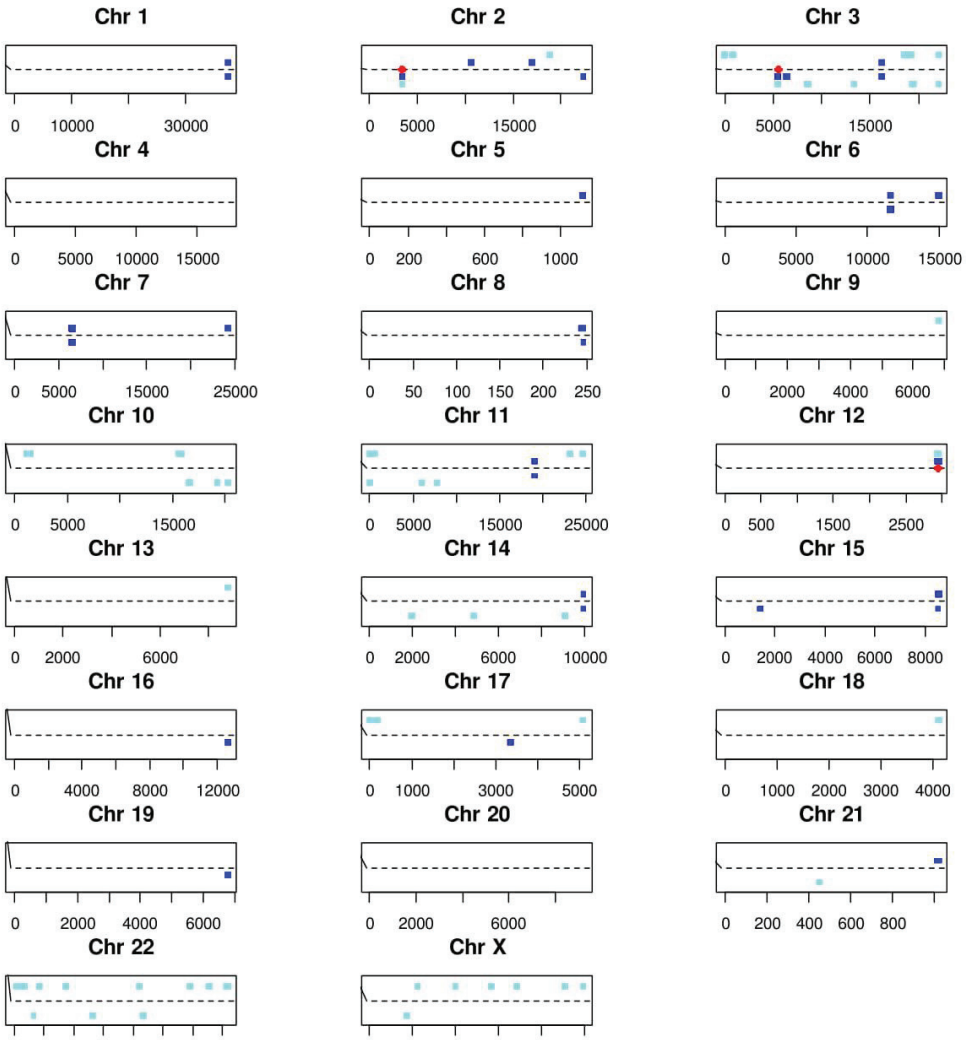
D) Alcohol intake: chr6:31650735-31651362

Additional file 6. DMRs and FL regions of folate in each chromosome.

Dark blue rectangles represent DMRs and light blue FL regions. Overlaps between the two methods are represented by red points.

Positive coefficients of the two methods are represented on the top part of each graphic and negative coefficients are on the bottom part. Positive (negative) coefficients of DMRs were set to 0.5 (-0.5) and positive (negative) coefficients of FL regions were set to 1 (-1) to clearly differentiate DMRs from FL regions.

x-axis represents the rank of CpG sites according to their position on the chromosome.

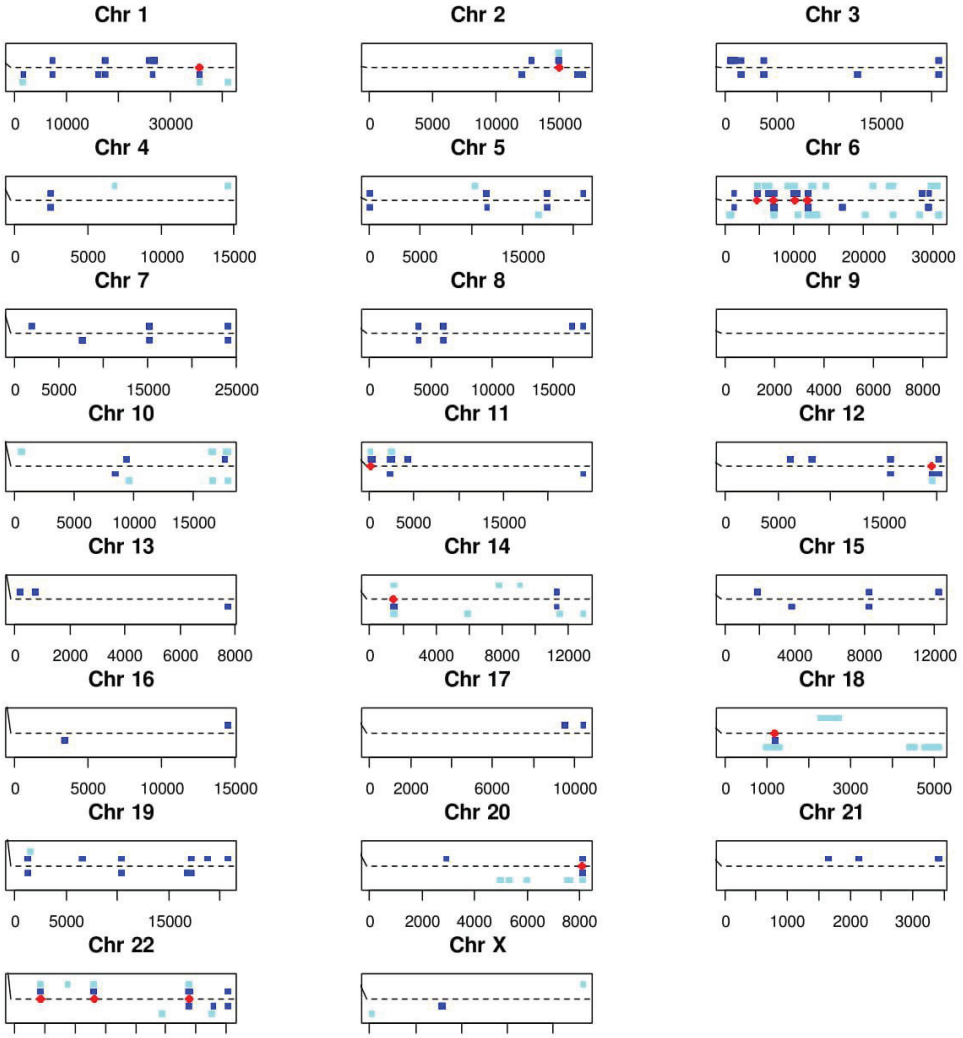


Additional file 7. DMRs and FL regions of alcohol in each chromosome.

Dark blue rectangles represent DMRs and light blue FL regions. Overlaps between the two methods are represented by red points.

Positive coefficients of the two methods are represented on the top part of each graphic and negative coefficients are on the bottom part. Positive (negative) coefficients of DMRs were set to 0.5 (-0.5) and positive (negative) coefficients of FL regions were set to 1 (-1) to clearly differentiate DMRs from FL regions.

x-axis represents the rank of CpG sites according to their position on the chromosome.



3- Average methylation and breast cancer risk

Context

A vast majority of human malignancies are associated with ageing, and age is a strong predictor of cancer risk. In particular BC is an age-associated disease whose incidence rises sharply after menopause (85). This increased risk was hypothesized to be the consequence of accumulation of genetic mutations associated with deregulation of cellular processes and genomic instability. Recently, DNA methylation-based marker of ageing, known as 'epigenetic clock' or 'DNA methylation age', can be used to accurately estimate the chronological age of all tissues and cell types (86). This composite biomarker of ageing, defined as a weighted average across 353 specific CpG sites, has been linked with several diseases including Alzheimer and Parkinson diseases (87, 88).

Departures of methylation-estimated age from chronological age can be used to define intrinsic epigenetic age acceleration (IEAA) that measures cell-intrinsic ageing effects, which are independent of chronological age and blood cell composition. IEAA has been used to predict lung cancer risk in a recent study (89). However, it is not yet known whether IEAA lends itself for predicting BC susceptibility in a prospective case-control study.

Objectives

- To evaluate whether intrinsic epigenetic age acceleration is associated with BC risk susceptibility.
- To investigate the association between global methylation and BC risk.

Approach

DNA methylation changes in 421,583 sites were profiled in 902 samples of a case-control study on BC nested within the EPIC cohort using the Illumina HumanMethylation 450K BeadChip arrays. One control participant was randomly assigned for each case based on: recruitment centre, length of follow-up, age at blood collection, time of blood collection, fasting status, menopausal status, menstrual cycle day and current use of contraceptive pill/hormone replacement therapy.

Overall global DNA methylation, defined as the mean methylation in the 421,583 sites, was computed for all participants. Global DNA methylation on specific regions of CpG sites reflecting their physical location in relation to CpG islands or based on a functional criterion was also performed. A conditional logistic regression was used to estimate global methylation (overall and for each category) association with BC risk.

The Horvath age estimation method (86) was used to calculate epigenetic age for each samples, based on the methylation levels of 353 CpG sites. Intrinsic epigenetic age acceleration (IEAA) was estimated as the residuals from a linear regression where epigenetic age was regressed on chronological age, adjusted for blood cells counts. Logistic regression was used to assess IEAA association with BC risk adjusted for known BC risks, such as alcohol consumption, BMI, age at menarche and physical activity. Stratification by menopausal status was also performed.

Main findings

Overall global DNA methylation was not significantly associated with BC risk whereas global DNA methylation in CpG islands was positively associated with BC risk (OR_{1SD}=1.20, $CI_{95\%}$=[1.03-1.40], p=0.02).

One unit increase in IEAA was significantly associated with a 4% increased risk of developing BC (OR=1.04; $CI_{95\%}$=[1.007-1.076]) in univariate analysis. Stratified analysis based on menopausal status revealed that IEAA was positively associated with development of postmenopausal BC (OR=1.06, $CI_{95\%}$=[1.019-1.110], p=0.003). The results were not attenuated after adjusting for known BC factors.

Conclusion

Assessed in blood samples, global methylation in CpG island regions and epigenetic age acceleration had a weak, but statistically significant, positive association with BC susceptibility. With an increased BC risk of 6% by one unit increase of IEAA and a p-value at 0.003, the association between epigenetic age acceleration and BC was more significant for postmenopausal women. Menopause has been known to accelerate age-related diseases including cancer. Age acceleration in postmenopausal BC may reflect differences in hormone exposure, which may explain why IEAA was only predictive of postmenopausal BC.

Published article: DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility.

European Journal of Cancer 75 (2017) 299–307

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.ejcancer.com

Original Research

DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility

Srikant Ambatipudi [a], Steve Horvath [b], Flavie Perrier [a], Cyrille Cuenin [a], Hector Hernandez-Vargas [a], Florence Le Calvez-Kelm [a], Geoffroy Durand [a], Graham Byrnes [a], Pietro Ferrari [a], Liacine Bouaoun [a], Athena Sklias [a], Véronique Chajes [a], Kim Overvad [c], Gianluca Severi [d,e,f], Laura Baglietto [d,f], Françoise Clavel-Chapelon [d], Rudolf Kaaks [g], Myrto Barrdahl [g], Heiner Boeing [h], Antonia Trichopoulou [i,j], Pagona Lagiou [i,j,k], Androniki Naska [i,j], Giovanna Masala [l], Claudia Agnoli [m], Silvia Polidoro [e], Rosario Tumino [n], Salvatore Panico [o], Martijn Dollé [p], Petra H.M. Peeters [q,r], N. Charlotte Onland-Moret [q], Torkjel M. Sandanger [s], Therese H. Nøst [s], Elisabete Weiderpass Vainio [s,t,u,v], J. Ramón Quirós [w], Antonio Agudo [x], Miguel Rodriguez-Barranco [y,z], José María Huerta Castaño [aa,z], Aurelio Barricarte [ab,ac,z], Ander Matheu Fernández [ad,ae], Ruth C. Travis [af], Paolo Vineis [ag], David C. Muller [ag], Elio Riboli [ag], Marc Gunter [a], Isabelle Romieu [a], Zdenko Herceg [a,*]

[a] *International Agency for Research on Cancer (IARC), Lyon, France*
[b] *Human Genetics and Biostatistics, University of California Los Angeles, Los Angeles, CA 90095-7088, USA*
[c] *Section for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark*
[d] *Inserm, Centre de Recherche en Epidémiologie et Santé des Populations (CESP, U1018), Université Paris-Saclay, Université Paris-Sud, UVSQ, Institut Gustave Roussy, Villejuif, France*
[e] *Human Genetics Foundation (HuGeF), Torino, Italy*
[f] *Cancer Epidemiology Centre, Cancer Council Victoria and Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourn, Australia*
[g] *Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany*
[h] *Department of Epidemiology, German Institute of Human Nutrition Potsdam-Rehbrücke, Nuthetal, Germany*
[i] *Hellenic Health Foundation, Athens, Greece*
[j] *WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, University of Athens Medical School, Athens, Greece*
[k] *Department of Epidemiology, Harvard School of Public Health, Boston, USA*

* *Corresponding author*: Epigenetics Group, International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, F-69008, Lyon, France. Fax: +33 4 72 73 83 22.
E-mail address: herceg@iarc.fr (Z. Herceg).

http://dx.doi.org/10.1016/j.ejca.2017.01.014

[l] Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy
[m] Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale Tumori, Milano, Italy
[n] Cancer Registry and Histopathology Unit, "Civic M.P. Arezzo" Hospital, ASP Ragusa, Italy
[o] Dipartimento di Medicina Clinica e Chirurgia, Federico II University, Naples, Italy
[p] Centre for Health Protection, National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands
[q] Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands
[r] MRC-PHE Centre for Environment and Health, Dept of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK
[s] Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway
[t] Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway
[u] Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
[v] Genetic Epidemiology Group, Folkhälsan Research Center, Helsinki, Finland
[w] Public Health Directorate, Asturias, Spain
[x] Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Catalan Institute of Oncology-IDIBELL, L'Hospitalet de Llobregat, Barcelona, Spain
[y] Escuela Andaluza de Salud Pública, Instituto de Investigación Biosanitaria ibsn Granada, Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain
[z] CIBER de Epidemiología y Salud Pública (CIBERESP), Spain
[aa] Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain
[ab] Navarra Public Health Institute, Pamplona, Spain
[ac] Navarra Institute for Health Research (IdiSNA) Pamplona, Spain
[ad] Cellular Oncology Group, Biodonostia Health Research Institute, Paseo Dr. Beguiristain s/n, San Sebastian, Spain
[ae] IKERBASQUE, Basque Foundation, Spain
[af] Cancer Epidemiology Unit, Nuffield Department of Population Health University of Oxford, Oxford UK
[ag] School of Public Health, Imperial College London, London, UK

Received 20 October 2016; received in revised form 16 December 2016; accepted 20 January 2017

KEYWORDS
DNA methylation;
Epigenomics;
Age acceleration;
Breast cancer;
Biomarkers;
Prospective studies

Abstract ***Aim of the study:*** A vast majority of human malignancies are associated with ageing, and age is a strong predictor of cancer risk. Recently, DNA methylation-based marker of ageing, known as 'epigenetic clock', has been linked with cancer risk factors. This study aimed to evaluate whether the epigenetic clock is associated with breast cancer risk susceptibility and to identify potential epigenetics-based biomarkers for risk stratification.
Methods: Here, we profiled DNA methylation changes in a nested case–control study embedded in the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort (n = 960) using the Illumina HumanMethylation 450K BeadChip arrays and used the Horvath age estimation method to calculate epigenetic age for these samples. Intrinsic epigenetic age acceleration (IEAA) was estimated as the residuals by regressing epigenetic age on chronological age.
Results: We observed an association between IEAA and breast cancer risk (OR, 1.04; 95% CI, 1.007–1.076, P = 0.016). One unit increase in IEAA was associated with a 4% increased odds of developing breast cancer (OR, 1.04; 95% CI, 1.007–1.076). Stratified analysis based on menopausal status revealed that IEAA was associated with development of postmenopausal breast cancers (OR, 1.07; 95% CI, 1.020–1.11, P = 0.003). In addition, methylome-wide analyses revealed that a higher mean DNA methylation at cytosine-phosphate-guanine (CpG) islands was associated with increased risk of breast cancer development (OR per 1 SD = 1.20; 95 %CI: 1.03–1.40, P = 0.02) whereas mean methylation levels at non-island CpGs were indistinguishable between cancer cases and controls.
Conclusion: Epigenetic age acceleration and CpG island methylation have a weak, but statistically significant, association with breast cancer susceptibility.

1. Introduction

Ageing is a major risk factor for most neoplasms [1]. In particular, breast cancer is an age-associated disease whose incidence rises sharply after menopause [1]. This increased risk was hypothesised to be the consequence of accumulation of genetic changes (mutations) associated with deregulation of cellular processes and genomic instability. However, accumulation of genetic changes exhibits striking interindividual differences [2], and differences in biological ageing processes may only be partly explained by genetic determinants [3].

A recent study demonstrates that DNA methylation (DNAm) data lend themselves for developing a highly accurate multitissue biomarker of ageing [4]. The DNAm-based marker of ageing (known as 'epigenetic clock') derived from several tissues can be used to accurately estimate the chronological age of all tissues and cell types [4]. This composite biomarker of ageing, which is defined as a weighted average across 353 specific CpG sites, produces an estimate of age (in units of years), referred to as 'epigenetic age' or 'DNA methylation age (DNAm age)'. Recent studies demonstrate that DNAm age is at least a passive biomarker of biological age: the epigenetic age of blood has been found to be predictive of all-cause mortality [5–9], frailty [10], cognitive and physical functioning [5]. Further, the utility of the epigenetic clock method using various tissues and organs has been demonstrated in applications surrounding Alzheimer disease [11], centenarian status [8], pre-natal and early life influences [12], Down syndrome [13], HIV infection [14], Huntington disease [15], obesity [16], lifetime stress [17], menopause [18], and Parkinson disease [19]. Departures of methylation-estimated age from chronological age can be used to define intrinsic epigenetic age acceleration (IEAA) that measures cell-intrinsic ageing effects that are independent of chronological age and blood cell composition.

A recent study suggests that IEAA can be used to predict lung cancer risk [20]. However, it is not yet known whether IEAA lends itself for predicting breast cancer susceptibility in a prospective case–control study. To test this hypothesis, we analysed blood methylation data from incident breast cancer cases and matching controls of a large prospective study within the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort.

2. Materials and methods

2.1. Selection of incident cancer and control participants

The present study was conducted on nested case–control samples from the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort, a large prospective study conducted in 23 centres across ten European countries (Denmark, France, Germany, Greece, Italy, Norway, Spain, Sweden, The Netherlands, and the United Kingdom), aiming to investigate the relationship between diet, lifestyle, metabolism and cancer risk [21]. In brief, the EPIC cohort includes a total of about 315,000 women and 200,000 men. At baseline recruitment, all study participants provided extensive questionnaire information about nutrition and other lifestyle factors. All study participants also provided a blood sample, which was processed, divided into aliquots of plasma, serum and buffy coat and frozen at −196 °C (under liquid nitrogen) for later use in specific research projects. In all EPIC centres, an identical protocol for subject recruitment, sample collection and storage was followed. Detailed information on the subject recruitment, baseline data, and blood collection protocols have been reported previously [22]. All participants gave written, informed consent for data and biospecimen collection and storage, as well as follow-up. The study was approved by the local ethics committees and the Institutional Review Board of the International Agency for Research on Cancer (IARC, Lyon, France). During prospective follow-up of the EPIC cohort, a very large number (>11,000) of newly diagnosed, invasive breast cancer cases were confirmed histologically or cytologically as primary breast cancers according to the International Classification of Diseases for Oncology, Second Edition (ICD-O-2) and included all breast cancer subsites (ICD C50.0-C50.9). A representative subset of these cases was used for studies comparing a variety of biomarker measurements with a set of control subjects, matching the cases by recruitment centre. Incident patients with cancer were identified at regular intervals through population-based cancer registries (in Denmark, Italy except Naples, the Netherlands, Norway, Spain, Sweden, and the United Kingdom) or by active follow-up (France, Germany, Greece, and Naples), which involved a combination of methods, including a review of health insurance records, cancer and pathology registries, and direct contact with participants and their next-of-kin.

For the purpose of this study, we included 960 females from the EPIC cohort including 480 incident breast cancer cases. Our main criteria for selection of case/control pairs included: (1) a balanced representation of the main subtypes of breast cancer, and (2) representation of recruiting centres. One control participant was randomly assigned for each patient with breast cancer from appropriate risk sets consisting of all cohort participants alive and free of cancer (except for non-melanoma skin cancer) at the time of diagnosis (and hence, age) of the index case. Matching criteria were: centre, length of follow-up, age at blood collection (3 months relaxed up to 2 years for sets without available controls), time of blood collection, fasting status, menopausal status, menstrual cycle day and current use of contraceptive pill/hormone replacement therapy.

Twenty technical replicates were included to compare inter- and intra-array batch variation. Technical replicates and 38 samples or their matched counterparts which failed the quality control criteria were excluded from the analysis leaving 902 participants (451 controls and 451 cases) (Table 1).

2.2. Bisulfite conversion and genome-wide DNA methylation analysis

The DNA was isolated as per the standard DNA extraction procedure from the from the buffy coat samples (Autopure LS, Qiagen). DNA methylome profiling was carried out using Illumina Infinium HumanMethylation450 (HM450) as previously described [23].

2.3. Bioinformatics analysis

Data preprocessing and analyses were performed using R 3.2.3 (https://www.r-project.org/) and Bioconductor 3.2 [24] as described before [23]. DNAm level was described as a β value, which is a continuous variable ranging between 0 (no methylation) and 1 (full methylation). To avoid spurious associations, we excluded the cross-reactive probes and probes overlapping with a known single nucleotide polymorphism (SNPs) with a minor allele frequency of at least 5% in the overall population (European ancestry, [25]), leaving 423,066 probes. In any given sample, probes with a detection P-value (a measure of an individual probe's performance) of more than 0.05 were assigned missing status. If a probe was missing in more than 5% of samples, it was excluded from all samples. According to this criterion, we excluded 1483 probes, leaving 421,583 probes available for the analyses. We applied colour bias correction followed by quantile and beta-mixture quantile normalisation (BMIQ) to align Type I and Type II probe distributions [26].

Table 1
Characteristics of incident breast cancer and control participants at baseline (i.e.time of blood collection).

	All samples	
	Controls (%)	Cases (%)
Sample size	451	451
Mean methylation (in %)	51.86	51.82
Age (years)		
Mean (SD)	52.3 (8.94)	52.3 (8.97)
Median	53.4	53.5
Alcohol consumption (g/d)		
Mean(SD)	8.2 (11.82)	10.0 (12.98)
Age at menarche		
Mean (SD)	12.9 (1.34)	12.7 (1.59)
BMI		
Mean (SD)	25.5 (4.22)	26.0 (4.72)
Physical activity (Cambridge index)		
Sedentary	99 (22.0)	121 (26.8)
Moderately sedentary	187 (41.5)	178 (39.5)
Moderately active	76 (16.9)	87 (19.3)
Active	78 (17.3)	62 (13.7)
Missing	11 (2.4)	3 (0.7)
Hormone receptor status		
$ER^+/PR^+/Her2^+$	–	85 (18.8)
$ER^+/PR^+/Her2^-$	–	290 (64.3)
$ER^-/PR^-/Her2^-$	–	76 (16.9)
Country		
Italy	160 (35.5)	160 (35.5)
Spain	27 (6.0)	27 (6.0)
UK	38 (8.4)	38 (8.4)
The Netherlands	66 (14.6)	66 (14.6)
Greece	25 (5.5)	25 (5.5)
Germany	135 (29.9)	135 (29.9)

SD: Standard deviation; ER: oestrogen receptor; PR: progesterone receptor; Her2: human epidermal growth factor receptor 2; BMI: body mass index.

2.4. White blood cell count estimates

Quantile normalised data were used to infer blood cell proportions. We estimate blood cell counts using two different software tools. First, Houseman's estimation method [27] was used to estimate the proportions of CD8+ T cells, CD4+ T, natural killer, B cells, and granulocytes (also known as polymorphonuclear leucocytes). Second, the advanced analysis option of the epigenetic clock software [4,14] was used to estimate the percentage of exhausted CD8+ T cells (defined as CD28-CD45RA-) and the number (count) of naïve CD8+ T cells (defined as CD45RA + CCR7+). We and others have shown that the estimated blood cell counts have moderately high correlations with corresponding flow cytometric measures [27,28]. For example, flow cytometric measurements correlate strongly with DNAm-based estimates: r = 0.63 for CD8+ T cells, r = 0.77 for CD4+ T cells, r = 0.67 for B cell, r = 0.68 for naïve CD8+ T cell, r = 0.86 for naïve CD4+ T, and r = 0.49 for exhausted CD8+ T cells [28].

2.5. Global and mean methylation analysis

For the global DNAm analyses, mean methylation of the DNAm probes (421,583) was calculated for cases and control samples. Human cancers are characterised by global hypomethylation and a loci-specific DNA hypermethylation [29]. We hypothesised that DNA methylation of probes would vary based on their physical location. To this end, the probes were classified into different categories either reflecting their physical location in relation to CpG islands (island, shore, shelf and open sea) or based on a functional criterion (DP: distal promoter, DS: distal sequence, GB: gene body, IG: intergenic, and PP: proximal promoter) as previously described [30]. A CpG shore is defined as the area 2 kb on either side of the CpG island, and a CpG shelf is defined as the area 2 kb outside of the CpG shore [31,32]. While the regions in the genome containing isolated CpG sites outside CpG islands, shores and

shelves, that do not have a specific designation are referred to as open seas [33].

2.6. Epigenetic clock of ageing

The epigenetic clock is a prediction method of chronological age based on the DNAm levels of 353 CpGs [4]. The predicted (estimated) age resulting from the epigenetic clock is referred to as 'DNA methylation age'. In IEAA, epigenetic age acceleration is defined as the DNAm age left unexplained by chronological age where intrinsic denotes a modification to this concept. In addition to adjusting for chronological age, IEAA also adjusts the DNAm age estimate for blood cell count estimates, arriving at a measure that is unaffected by both variation in chronological age and blood cell composition.

We focussed on IEAA in our blood-based methylation study as this measure of age acceleration is significantly correlated with epigenetic age acceleration in (non-malignant) female breast tissue [9].

Formally, IEAA is defined by regressing DNAm age on chronological age and seven measures of blood cell count abundances: naive CD8 T cells, exhausted CD8 T cells (defined as CD28-CD45RA-), plasma blasts, CD4 T cells, NK cells, monocytes, granulocytes. IEAA is automatically calculated using the advanced analysis option of the epigenetic clock software (where IEAA is denoted as 'AAHOAdjCellCounts'). A positive or negative value of IEAA indicates that the woman is either older or younger than expected based on chronological age at the time of the blood draw.

2.7. Statistical analysis

For the mean methylation analysis, average methylation over all probes within each category was calculated and the odds ratios (per one standard deviation of global methylation) were estimated by conditional logistic regression model with case–control status as the outcome and the epigenome-wide methylation measurement as continuous predictor adjusting for surrogate variables (technical batch effects such as sample plate, array chips), alcohol consumption (g/day) and body mass index (BMI) as continuous variable.

Odds ratios (ORs) for breast cancer and 95% CIs were calculated by using logistic regression for IEAA. Initial analysis was done using unconditional logistic regression to allow calculation of OR. Multivariate logistic regression was performed by including known breast cancer risk factors including alcohol consumption (g/day), full term pregnancy (ever/never), BMI (as continuous variable and as categorical variable: underweight, normal, overweight and obese), level of education (none, primary, technical/profession, secondary, higher education), age at menarche, Cambridge physical activity index (inactive, moderately inactive, moderately active and active) stratified by clustering variable. A stratified multivariate conditional logistic regression analysis based on the menopausal status was performed using the aforementioned models.

3. Results

3.1. Baseline characteristics

The baseline characteristics of samples at the time of recruitment are listed in Table 1. Women were between 26 and 73 years of age with a mean age of 52.3 years for cases and controls. The majority of breast cancer cases were hormone receptor (ER and PR) positive (83%) while 17% of the breast cancers were triple negative (Table 1). There was a very high correlation between the intra- and interplate technical replicates (average correlation coefficient $r^2 = 0.98$ and 0.97, respectively, data not shown).

3.2. Hypermethylation of CpG islands is associated with breast cancer risk

We compared the global mean methylation across 421,583 probes and observed no difference between prospectively collected cases and matched controls (51.82% versus 51.86%, $P = 0.68$). Our analysis showed that each unit (95% CI/1SD, 1.03–1.40, $P = 0.02$) increase in methylation at CpG island sites increased the risk of being a case by 20% (Table 2). While $P < 0.05$, it should be noted that the results would be marginally significant allowing for four subsets (CpG islands, CGI shores, CGI shelves, and open sea). No change in breast cancer risk was observed for other regions (shore, shelf and open sea) (Table 2), nor did we find an association of individual CpG site or region with breast cancer status.

Table 2
Association between global methylation and breast cancer risk by CpG genomic features.

	Context	# CpGs	Std. dev.	OR (95% CI)[a]	P value
	All CpG sites	421 583	3.45E-04	1.09 (0.94–1.25)	0.21
	Islands	130 982	5.87E-04	1.20 (1.03–1.40)	0.02
	Open Sea	150 852	4.50E-03	1.49 (0.36–6.24)	0.58
CpG	Shelf	40 948	4.88E-04	0.89 (0.78–1.02)	0.10
context	Shore	98 801	5.40E-04	1.00 (0.87–1.16)	0.97
	Distal promoter	19 990	5.42E-04	1.06 (0.92–1.21)	0.44
	Distal sequence	7828	6.68E-04	0.96 (0.84–1.09)	0.52
Genic	Gene Body	168 460	3.80E-04	1.02 (0.89–1.18)	0.76
context	Intergenic	56 903	5.35E-04	1.02 (0.89–1.17)	0.76
	Proximal promoter	168 337	5.26E-04	1.15 (0.99–1.34)	0.07

[a] Odds ratio and confidence interval were calculated per 1 standard deviation. Odds ratios were adjusted for body mass index (BMI) (continuous variable) and daily alcohol intake. OR- Odds ratio, CI: confidence interval.

3.3. Postmenopausal breast cancer cases exhibit DNA methylation age acceleration

Epigenetic age had a strong positive correlation with chronological age in both case and control samples (Fig. 1a). We observed a marginally significant difference in age acceleration between prospective cases compared to matched controls (Fig. 1b, $P = 0.05$, Supplementary Fig. 1). Stratified analysis based on time from blood collection to disease diagnosis revealed that prospective breast cancers exhibited age acceleration 10 years prior to diagnosis compared to matched control samples (Fig. 1c, $P = 0.01$).

A conditional logistic regression model that relates breast cancer status to IEAA showed that IEAA was associated (Table 3) with breast cancer status. The results were not attenuated after adjusting for known breast cancer factors (Supplementary Table 1). Each unit increase in IEAA led to 4% increased odds of being a breast cancer case (OR, 1.04; 95% CI, 1.007–1.076, $P = 0.016$) (Table 3). IEAA follows an approximately normal distribution with mean zero, variance = 28.2, standard deviation of 5.31. The following quantiles describe the empirical distribution of IEAA: minimum = −24.2, maximum 24.4, median = −0.12, first quartile = −3.0, third quartile = 3.0. Thus, 25% of women had an IEAA value > 3.

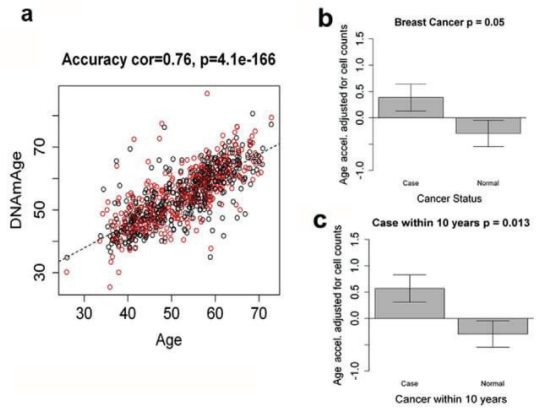


Fig. 1. Epigenetic clock analysis. a) DNA methylation age (y-axis) versus chronological age (x-axis). Points correspond to female subjects. Red indicates breast cancer case, black control. The dashed line indicates a regression line, b) epigenetic age acceleration versus breast cancer status. Each bar plot depicts the mean value, standard deviation and reports a non-parametric group comparison test p-value (Wilcoxon test), c) epigenetic age acceleration versus breast cancer status (developed within 10 years post blood draw). Each bar plot depicts the mean value, standard deviation and reports a non-parametric group comparison test p-value (Wilcoxon test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Logistic regression analysis of IEAA for incident breast cancer status.

	Univariate analysis OR (95% CI)	Multivariate analysis[a] OR (95% CI)
All samples		
IEAA	1.04 (1.007–1.075)	1.04 (1.007–1.076)
Premenopausal samples		
IEAA	1.00 (0.9572–1.06)	1.00 (0.9510–1.056)
Postmenopausal samples		
IEAA	1.06 (1.019–1.11)	1.07 (1.020–1.11)

OR: Odds Ratio; CI: Confidence Interval; IEAA: Intrinsic Epigenetic Age Acceleration.

[a] Odds ratios were adjusted for physical activity (inactive, moderately inactive, moderately active and active).

None of the blood cell count measures were associated with disease status in prediagnostic blood samples (Supplementary Fig. 2). Interestingly, high physical activity was associated with decreased odds of being a breast cancer case (Supplementary Table 1).

A recent study demonstrated that menopause has a weak but statistically significant effect on epigenetic age acceleration. Further, menopause has been known to accelerate age-related diseases including breast cancer [34,35]. To adjust for menopausal status, we evaluated the association between IEAA and breast cancer in separate strata defined by menopausal status (premenopausal and postmenopausal). The baseline characteristics of premenopausal and postmenopausal breast samples are shown in Supplementary Table 2. We observed a positive correlation between epigenetic and chronological age in postmenopausal samples (Fig. 2a). Stratified analysis of postmenopausal breast cancers based on the lead-time between blood collection and cancer diagnosis revealed that breast cancers had a higher IEAA compared to non-cancer samples (Fig. 2b, Supplementary Fig. 3).

A very high value of IEAA = 10 is associated with a doubling of odds of developing postmenopausal breast cancer (OR = 1.97 (1.22–2.83) calculated as 1.07^{10} from our multivariate logistic regression model Table 3). Twenty-five percent of all women exhibit an age acceleration larger than 3 which is associated with 22% increase in the odds of developing postmenopausal breast cancer (OR = 1.22 (1.06–1.37) calculated as 1.07^{3}).

We found that breast cancer that developed within 10 years from date of recruitment had a stronger association with IEAA (Fig. 2c). However, the results of this secondary analysis should be interpreted with caution due to an inflated false positive rate resulting from multiple comparisons. We did not observe such associations in premenopausal breast samples (Supplementary Figs. 4, 5). Similar to our findings in all breast samples, high physical activity was associated with decreased odds of being a breast cancer case in postmenopausal women (Supplementary Table 3).

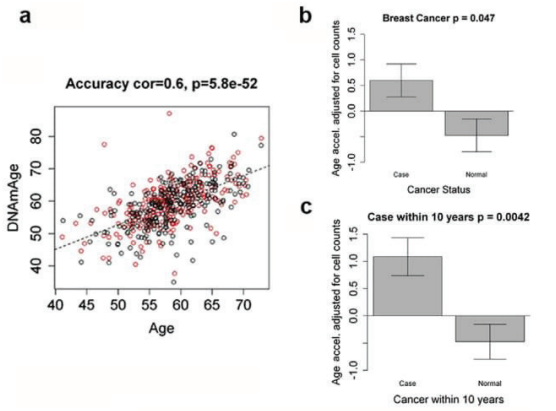


Fig. 2. Epigenetic clock analysis for postmenopausal samples. a) DNA methylation age (y-axis) versus chronological age (x-axis). Points correspond to female subjects. Red indicates breast cancer case, black control. The dashed line indicates a regression line; b) epigenetic age acceleration versus breast cancer status. Each bar plot depicts the mean value, standard deviation and reports a non-parametric group comparison test p-value (Wilcoxon test); c) epigenetic age acceleration versus breast cancer status (developed within 10 years post blood draw). Each bar plot depicts the mean value, standard deviation, and reports a non-parametric group comparison test p-value (Wilcoxon test). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Interestingly, we observed a highly significant association between IEAA and incident postmenopausal breast cancers (OR, 1.07; 95% CI, 1.020–1.11, $P = 0.003$). By contrast, no significant association could be observed for incident premenopausal breast cancers (OR, 1.00; 95% CI, 0.9510–1.056, $P = 0.94$) (Table 3).

4. Discussion

Using a rigorous and large-scale nested prospective case–control study, we demonstrate that: (1) IEAA in blood increases the odds of developing postmenopausal breast cancers and (2) genome-wide hypermethylation in CpG islands is associated with incident breast cancer cases. While several articles have studied blood methylation data versus breast cancer risk [36–39], it appears that ours is the first study to detect a weak but significant association of IEAA with breast cancer susceptibility. Our study stands out in terms of its large sample size, its use of a robust epigenome wide technology (Illumina 450K array), the careful matching of breast cancer cases with controls in a prospective case–control study, and its use of a powerful epigenetic biomarker of ageing, which is independent of blood cell counts (IEAA).

Our finding regarding the association between global CpG island methylation levels and breast cancer risk is congruent with the findings from our earlier retrospective study on breast cancer [39] and supports the notion that regulatory regions of the genome are often hypermethylated in cancer cells [29]. It is noteworthy that we observed CpG island hypermethylation in blood tissue samples of incident breast cancer patients. Several epidemiological case–control studies have reported global genomic hypomethylation in peripheral blood of cancer patients, suggesting a systemic effect of hypomethylation on disease predisposition [40,41]. In addition, two recent studies reported a lower global methylation levels in prospectively collected blood samples from breast cancer cases compared to controls [38,42]. However, we did not find any change in global DNAm levels between cases and controls. These discrepancies may be due to technical and biological variations attributable to the low power of the studies.

Epigenetic changes are ubiquitous in primary breast cancers although the role of deregulation of the epigenome is largely unknown. It has been suggested that a gradual accumulation of methylation changes ('epigenetic drift') may occur through stochastic events, resulting in clonal expansion of the stem/progenitor cells, and that this process may contribute to the age-associated increase in risk of developing breast cancer [43–45]. DNAm age is highly correlated to chronological age across sorted cell types (CD4 T cells, monocytes, B cells, glial cells, neurons), complex tissues (e.g. blood) and organs (brain, breast, kidney, liver, lung) [4]. Our findings were consistent with the previous studies in different tissues [4,16]. The epigenetic clock derived from the DNAm age is robust with respect to the batch effects and can be applied to all Illumina array platforms: the EPIC chip (850K), the Illumina 450K array and the 27K array [4] and possibly measures a cell intrinsic and tissue independent epigenetic drift [46]. For blood derived DNA measured on the Illumina 450K array, the epigenetic clock algorithm provides not only several measures of age acceleration but also estimates of blood cell counts. One of the major concerns regarding age-associated DNAm signatures is the influence of tissue's cellular composition which may alter with age. We found no differences in leucocyte subpopulations between cases and controls. By definition, our intrinsic measure of epigenetic age acceleration (IEAA) is not confounded by changes in the proportion of blood cell counts (Methods). We focussed on IEAA as it has been shown to be correlated with epigenetic age acceleration in breast tissue [9]. Future research could investigate whether epigenetic age acceleration of breast tissue is predictive of breast cancer.

We can only speculate when it comes to explaining why IEAA was only predictive of postmenopausal breast cancer but not of premenopausal breast cancer. Breast cancers developing in postmenopausal women are influenced by specific polymorphisms in endogenous steroid hormone metabolic pathways and exogenous administration of hormones at menopause (hormone

replacement therapy). Our observed age acceleration in postmenopausal breast cancers might reflect differences in hormone exposure. In this context, it is noteworthy that both natural and surgical menopause are associated with an increase in intrinsic age acceleration [18]. In addition, age-associated compromised detoxification, DNA repair mechanisms and immune surveillance may add to the endogenous factors which could lead to postmenopausal breast cancer development [1]. It is unlikely that smoking and BMI confound the relationship between epigenetic age and breast cancer risk because : (1) BMI and smoking have only a very weak effect on the epigenetic age acceleration of blood tissue (correlation $r < 0.10$) [16,20], and (2) we could detect accelerated ageing effects in multivariate regression models that adjusted for these potential confounders. Our results based on a prospective study cohort points to a higher rate of ageing in the blood samples from individuals who develop breast cancer compared to the controls. While the results from our epigenetic age analysis are biologically meaningful, the association between DNAm age and disease risk is probably too weak for prognostic purposes.

In the present study, we demonstrated that a surrogate tissue (blood) captures accelerated ageing effects and relates to an effector (breast cancer) of ageing. We have demonstrated that IEAA was associated with postmenopausal breast cancer susceptibility and identified potential epigenetics-based biomarkers for risk stratification. Because menopause has been known to accelerate age-related diseases including cancer, our finding also suggest potential underlying mechanism and provides biological plausibility to the association between menopause and cancer risk. Further research aimed at understanding epigenome deregulation in cancer causation, risk stratification and the mechanism underlying accelerated epigenetic clock is warranted.

Role of funding resource

The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Conflict of interest statement

The Regents of the University of California is the sole owner of a patent application directed at the invention of measures of epigenetic age acceleration for which Steve Horvath is a named inventor. The other authors declare no conflict of interest.

Funding

This work was supported by grants from the Institut National du Cancer (INCa, France) (2012-070) to IR and ZH and the European Commission (EC) Seventh Framework Programme (FP7 Exposomics) (308610-2) Translational Cancer Research (TRANSCAN) (TRANS201301184) Framework and the Fondation Association pour la Recherche contre le Cancer (ARC, France) to ZH. ZH was also supported by the EC FP7 EurocanPlatform: A European Platform for Translational Cancer Research (grant number: 260791). Funding for the work also comes from a grant from: "Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy" and the Regional Government of Asturias. EPIC Greece is supported by the Hellenic Health Foundation. SH was supported by NIH/NIA 1U34AG051425-01. The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Acknowledgements

The work reported in this article was undertaken during the tenure of a Postdoctoral Fellowship (to SA) from the International Agency for Research on Cancer, partially supported by the EC FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND). A.S. is supported by the PhD fellowship from the Fonds National de la Recherche, Luxembourg (AFR Code: 10100060). This study depended on the participation of the women in the EPIC cohort, to whom we are grateful.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ejca.2017.01.014.

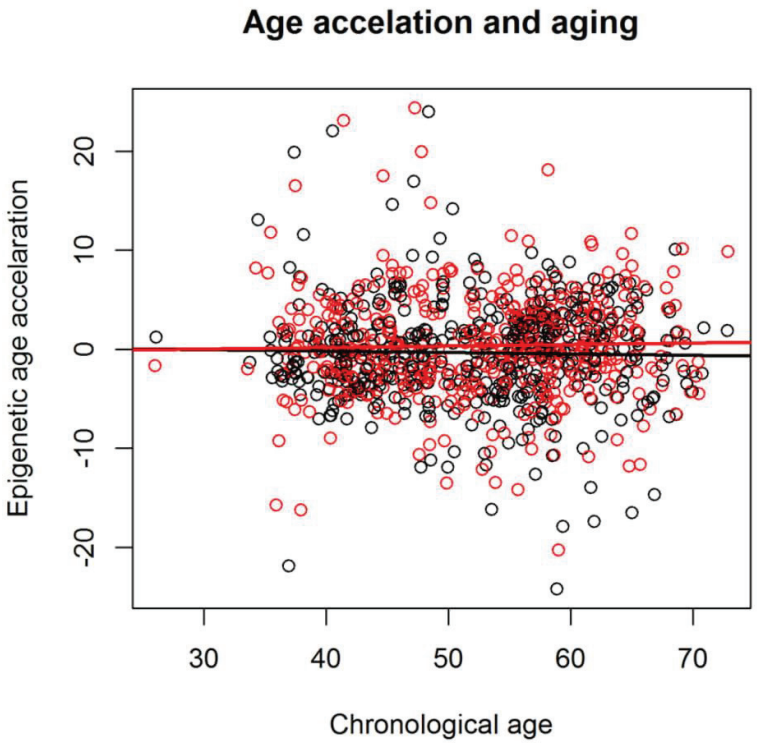
References

[1] Benz CC. Impact of aging on the biology of breast cancer. Crit Rev Oncol Hematol 2008;66(1):65–74.
[2] Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science 2015;348(6237):880–6.
[3] Deelen J, Beekman M, Capri M, Franceschi C, Slagboom PE. Identifying the genomic determinants of aging and longevity in human population studies: progress and challenges. Bioessays 2013;35(4):386–96.
[4] Horvath S. DNA methylation age of human tissues and cell types. Genome Biol 2013;14(10):R115.
[5] 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.
[6] Christiansen L, Lenart A, Tan Q, Vaupel JW, Aviv A, McGue M, et al. DNA methylation age is associated with mortality in a longitudinal Danish twin study. Aging Cell 2016;15(1):149–54.
[7] Perna L, Zhang Y, Mons U, Holleczek B, Saum KU, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clin Epigenetics 2016;8:64.

[8] Horvath S, Pirazzini C, Bacalini MG, Gentilini D, Di Blasio AM, Delledonne M, et al. Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. Aging (Albany NY) 2015;7(12):1159–70.
[9] 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.
[10] Breitling LP, Saum KU, Perna L, Schottker B, Holleczek B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. Clin Epigenetics 2016;8:21.
[11] Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. Aging (Albany NY) 2015;7(12):1198–211.
[12] Simpkin AJ, Hemani G, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal and early life influences on epigenetic age in children: a study of mother–offspring pairs from two cohort studies. Hum Mol Genet 2016;25(1):191–201.
[13] Horvath S, Garagnani P, Bacalini MG, Pirazzini C, Salvioli S, Gentilini D, et al. Accelerated epigenetic aging in Down syndrome. Aging Cell 2015;14(3):491–5.
[14] Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis 2015;212(10):1563–73.
[15] Horvath S, Langfelder P, Kwak S, Aaronson J, Rosinski J, Vogt TF, et al. Huntington's disease accelerates epigenetic aging of human brain and disrupts DNA methylation levels. Aging (Albany NY) 2016;8(7):1485–512.
[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.
[17] Zannas AS, Arloth J, Carrillo-Roa T, Iurato S, Roh S, Ressler KJ, et al. Lifetime stress accelerates epigenetic aging in an urban, African American cohort: relevance of glucocorticoid signaling. Genome Biol 2015;16:266.
[18] Levine ME, Lu AT, Chen BH, Hernandez DG, Singleton AB, Ferrucci L, et al. Menopause accelerates biological aging. Proc Natl Acad Sci U S A 2016;113(33):9327–32.
[19] Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging (Albany NY) 2015;7(12):1130–42.
[20] Levine ME, Hosgood HD, Chen B, Absher D, Assimes T, Horvath S. DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. Aging (Albany NY) 2015;7(9):690–700.
[21] Bingham S, Riboli E. Diet and cancer–the European prospective investigation into cancer and nutrition. Nat Rev Cancer 2004;4(3):206–15.
[22] Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European prospective investigation into cancer and nutrition (EPIC): study populations and data collection. Public Health Nutr 2002;5(6B):1113–24.
[23] Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Calvez-Kelm FL, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics 2016;8(5):599–618.
[24] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.
[25] Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics 2013;8(2):203–9.
[26] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 2013;29(2):189–96.
[27] Houseman AE, Accomando PW, Koestler CD, Christensen CB, Marsit JC, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinforma 2012:13.
[28] 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.
[29] Esteller M. Epigenetics in cancer. N Engl J Med 2008;358(11):1148–59.
[30] Martin M, Ancey PB, Cros MP, Durand G, Le Calvez-Kelm F, Hernandez-Vargas H, et al. Dynamic imbalance between cancer cell subpopulations induced by transforming growth factor beta (TGF-beta) is associated with a DNA methylome switch. BMC Genomics 2014;15:435.
[31] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics 2011;98(4):288–95.
[32] Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet 2009;41(2):178–86.
[33] Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics 2011;6(6):692–702.
[34] Blagosklonny MV. Why men age faster but reproduce longer than women: mTOR and evolutionary perspectives. Aging (Albany NY) 2010;2(5):265–73.
[35] Horiuchi S. Postmenopausal acceleration of age-related mortality increase. J Gerontol A Biol Sci Med Sci 1997;52(1):B78–92.
[36] Demetriou CA, Chen J, Polidoro S, van Veldhoven K, Cuenin C, Campanella G, et al. Methylome analysis and epigenetic changes associated with menarcheal age. PLoS One 2013;8(11):e79391.
[37] Heyn H, Carmona FJ, Gomez A, Ferreira HJ, Bell JT, Sayols S, et al. DNA methylation profiling in breast cancer discordant identical twins identifies DOK7 as novel epigenetic biomarker. Carcinogenesis 2013;34(1):102–8.
[38] Severi G, Southey MC, English DR, Jung CH, Lonie A, McLean C, et al. Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. Breast Cancer Res Treat 2014;148(3):665–73.
[39] Xu X, Gammon DM, Hernandez-Vargas H, Herceg Z, Wetmur GJ, Teitelbaum LS, et al. DNA methylation in peripheral blood measured by LUMA is associated with breast cancer in a population-based study. FASEB J 2012;26:2657–66.
[40] Wallace K, Grau MV, Levine AJ, Shen L, Hamdan R, Chen X, et al. Association between folate levels and CpG Island hypermethylation in normal colorectal mucosa. Cancer Prev Res (Phila) 2010;3(12):1552–64.
[41] Kuchiba A, Iwasaki M, Ono H, Kasuga Y, Yokoyama S, Onuma H, et al. Global methylation levels in peripheral blood leukocyte DNA by LUMA and breast cancer: a case-control study in Japanese women. Br J Cancer 2014;110(11):2765–71.
[42] 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(1):67.
[43] Issa JP. Aging, DNA methylation and cancer. Crit Rev Oncol Hematol 1999;32(1):31–43.
[44] Langevin SM, Pinney SM, Leung YK, Ho SM. Does epigenetic drift contribute to age-related increases in breast cancer risk? Epigenomics 2014;6(4):367–9.
[45] Issa JP. Aging and epigenetic drift: a vicious cycle. J Clin Invest 2014;124(1):24–9.
[46] Zheng Y, Joyce BT, Colicino E, Liu L, Zhang W, Dai Q, et al. Blood epigenetic age may predict cancer incidence and mortality. EBioMedicine 2016;5:68–73.

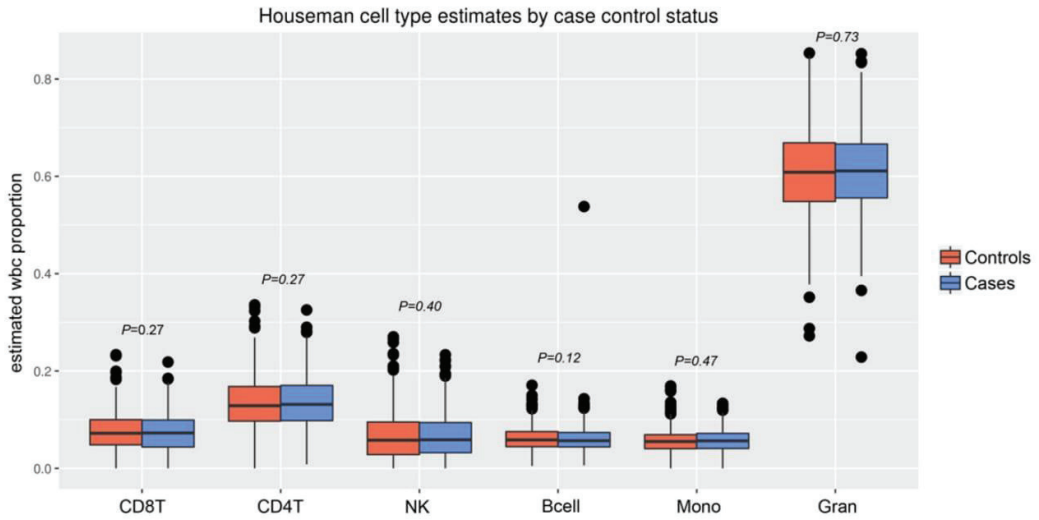
Supplementary Figure (online publication only)

Supplementary Material

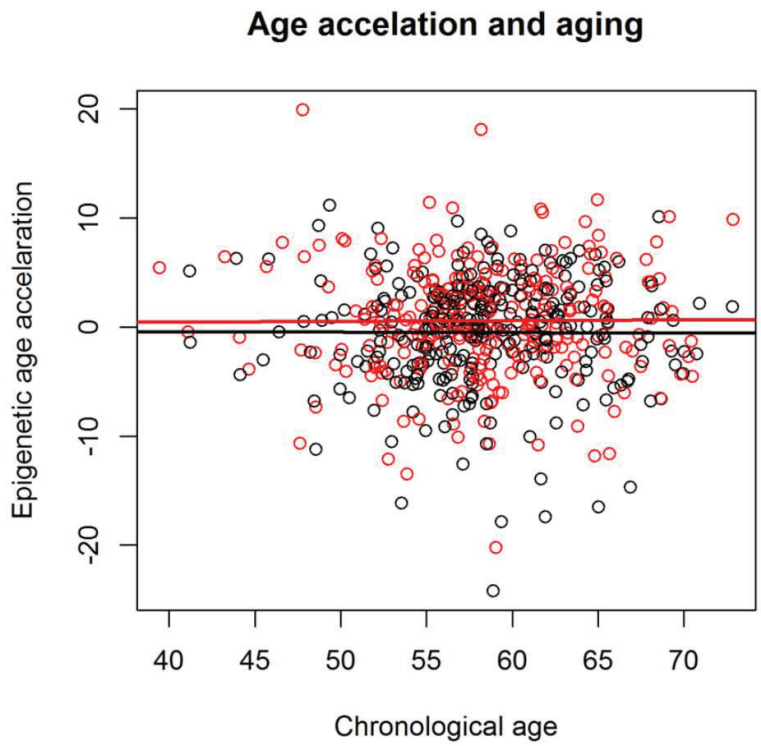


Supplementary Figure 1: Epigenetic age accelaration of breast samples.

Epigenetic age accelaration (IEAA) (Y-aixs) versus chronological age. Points correspond to female subjects. Red colored circles indicates breast cancer case while the black circles represent non-case samples. The solid lines indicates a regression lines for cases (in red) and non-case samples (in black).

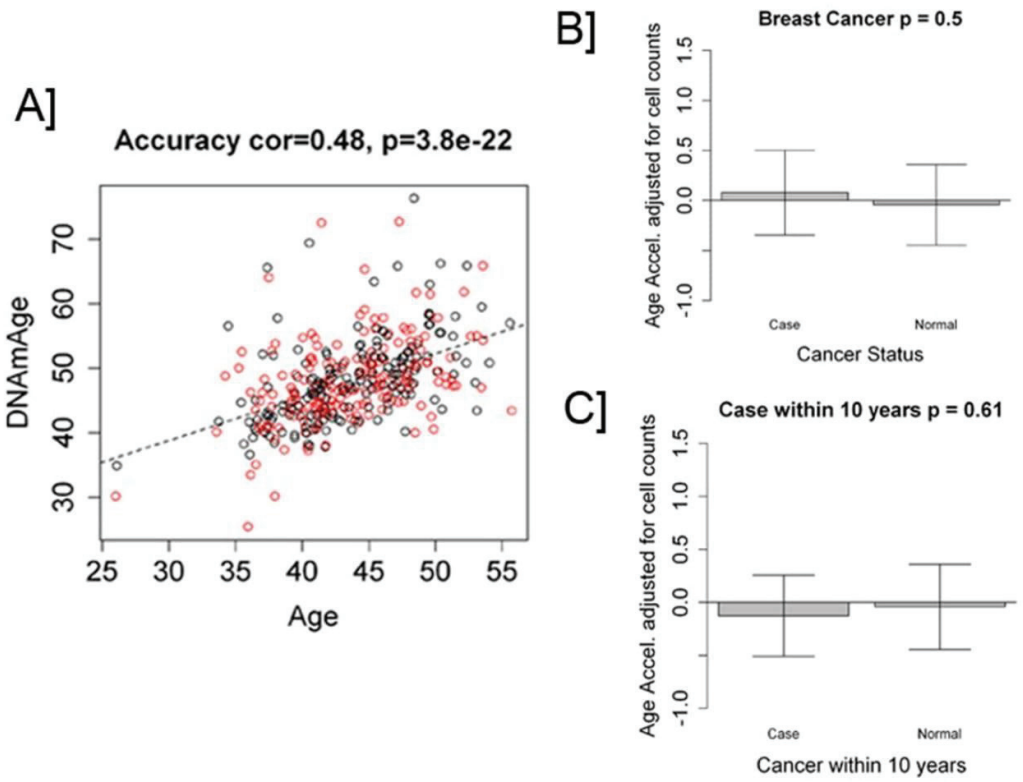


Supplementary Figure 2: Distribution of inferred leucocyte cell subpopulation. Proportion of leukocyte subtypes derived from DNA methylation data. Inferred data were plotted by sample groups (breast cancer cases and controls) where X-axis shows leucocyte subtypes and Y-axis shows proportion of estimated leucocytes.

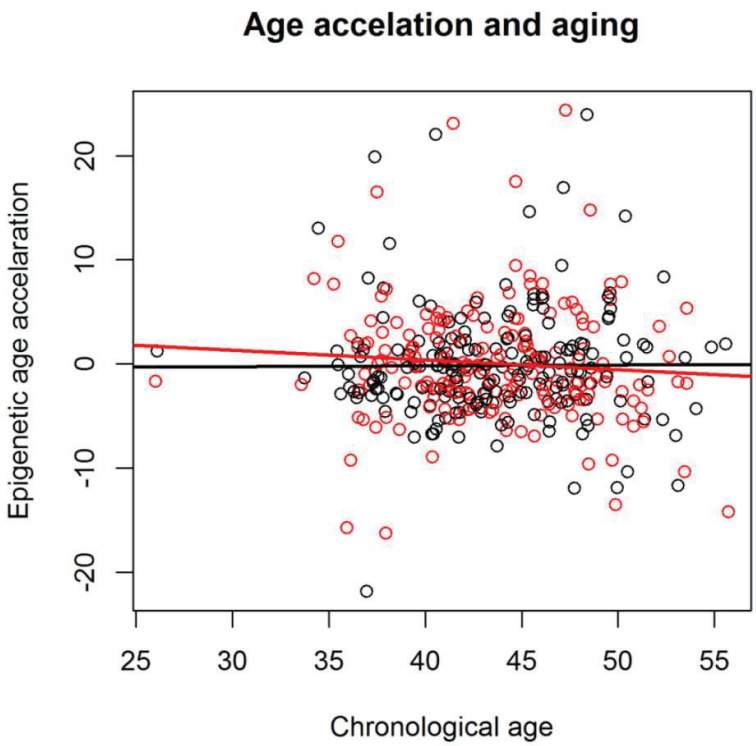


Supplementary Figure 3: Epigenetic age accelaration of postmenopausal breast samples.

Epigenetic age accelaration (IEAA) (Y-aixs) versus chronological age. Points correspond to female subjects. Red colored circles indicates breast cancer case while the black circles represent non-case samples. The solid lines indicates a regression lines for cases (in red) and non-case samples (in black)..



Supplementary Figure 4: Epigenetic clock analysis for premenopausal breast samples. A) DNAm age (Y-aixs) versus chronological age. Points correspond to female subjects. Red colored circles indicates breast cancer case while the black circles represent non-case samples. The dashed line indicates a regression line. B) Epigenetic age accelaration versus breast cancer status. Each bar plot depicts the mean value, standard deviation, and reports a non-parametric group test p- value (Wilcoxon test). C) Epigenetic age accelaration versus breast cancer status (developed within 10 years post blood draw). Each bar plot depicts the mean value, standard deviation, and reports a non-parametric group comparison test p- value (Wilcoxon test).



Supplementary Figure 5: Epigenetic age accelaration of premenopausal breast samples. Epigenetic age accelaration (IEAA) (Y-aixs) versus chronological age. Points correspond to female subjects. Red colored circles indicates breast cancer case while the black circles represent non-case samples. The solid lines indicates a regression lines for cases (in red) and non-case samples(in black).

Supplementary Table 1. Conditional logistic regression model of epigenetic age acceleration in all samples

	OR (95% CI)	P value
IEAA	1.05 (1.01-1.09)	0.01
Alcohol at the time of recruitment	1.01 (1.00-1.02)	0.06
Level of education (Ref. No education)		
Primary	1.53 (0.63-3.69)	0.34
Technical/professional	1.14 (0.44-2.97)	0.79
Secondary	1.92 (0.72-5.08)	0.19
Higher education	1.69 (0.63-4.51)	0.29
Full term pregnancy (Ever/never)	0.96 (0.66-1.41)	0.85
Physical activity (Cambridge index, Ref.		
Moderately inactive	0.71 (0.48-1.06)	0.10
Moderately active	0.77 (0.48-1.24)	0.29
Active	0.54 (0.33-0.91)	0.02
Age at menarche	0.93 (0.85-1.02)	0.14
BMI (Categorical, Ref. Normal)		
Underweight	0.61 (0.14-2.68)	0.51
Overweight	1.01 (0.72-1.42)	0.95
Obese	1.15 (0.75-1.78)	0.52

Conditional logistic regression was performed using known breast cancer risk factors highlighted in bold

IEAA: Intrinsic Epigenetic Age Acceleration

BMI: Body Mass Index

Supplementary Table 2. Demographic and lifestyle factor details of pre and postmenopausal samples

		Premenopausal samples		Postmenopausal samples	
		Controls (%)	Cases (%)	Controls (%)	Cases (%)
Sample size		180	180	259	259
		Demographic and lifestyle factors			
Age (years)					
	Mean (SD)	43.6 (4.73)	43.6 (4.74)	58.5 (5.50)	58.5 (5.50)
	Median	43.5	43.4	58.3	58.3
Smoking					
	Never	90 (50.0%)	85 (47.3%)	158 (61.0%)	171 (66.1%)
	Former	37 (20.5%)	46 (25.5%)	50 (19.3%)	47 (18.1%)
	Current	51 (28.4%)	49 (27.2%)	50 (19.3%)	40 (15.4%)
	Not known	2 (1.1%)	-	1 (0.4%)	1 (0.4%)
Alcohol					
	Mean(SD)	8.1(11.06)	10.3 (12.12)	8.1 (12.15)	9.5 (13.55)
	Median	4.4	5.3	3.0	4.0
Age at					
	Mean (SD)	12.9 (1.34)	12.7 (1.59)	13.3 (1.64)	13.3 (1.71)
	Median	13.0	13.0	13.0	13.0
BMI					
	Mean (SD)	24.7 (4.14)	24.8 (4.12)	26.1 (4.25)	26.9 (4.95)
	Median	23.88	23.98	25.56	25.97
IEAA					
	Mean (SD)	-0.042 (5.39)	0.079 (5.67)	-0.47 (5.16)	0.60 (5.19)

IEAA: Intrinsic Epigenetic Age Acceleration

Supplementary Table 3: Conditional logistic regression model of epigenetic age acceleration in postmenopausal samples

	OR (95% CI)	P value
IEAA	1.08 (1.03-1.13)	0.003
Alcohol at the time of recruitment	1.01 (0.99-1.02)	0.424
Level of education (Ref. No education)		
Primary	2.94 (0.75-11.46)	0.121
Technical/professional	1.46 (0.34-6.20)	0.609
Secondary	2.51 (0.57-11.13)	0.226
Higher education	2.98 (0.67-13.20)	0.151
Full term pregnancy (Ever/never)	0.94 (0.56-1.58)	0.827
Physical activity (Cambridge index, Ref.		
Moderately inactive	0.78 (0.47-1.29)	0.334
Moderately active	0.51 (0.26-0.99)	0.046
Active	0.39 (0.19-0.80)	0.011
Age at menarche	0.98 (0.86-1.11)	0.759
BMI (Categorical, Ref. Normal)		
Underweight	0.61 (0.05-7.77)	0.707
Overweight	1.11 (0.71-1.74)	0.653
Obese	1.08 (0.62-1.88)	0.791

Conditional logistic regression was performed using known breast cancer risk factors highlighted in bold

IEAA: Intrinsic Epigenetic Age Acceleration

BMI: Body Mass Index

CONCLUSION AND PERSPECTIVES

Conclusion

Recent technical progress in the acquisition of biological features generated an exponential growth of the amount of data expressing a wealth of biological parameters. For example, methylation levels of 27K CpG sites were first measured in 2008 via the Illumina Infinium HumanMethylation27 BeadChip platform. Then in 2011, a new array was developed to target over 450K CpG sites. Currently, since 2015, the Illumina MethylationEPIC BeadChip array covers over 850K sites. We can easily imagine that these numbers will further increase in the near future. To analyze high-throughput datasets, there will be progressively demanding needs of statistical tools to fully exploit the potential of these data.

In this thesis, the statistical tools used for the analysis of epigenetics data were instrumental to handle the high dimensionality and complexity of DNA methylation data. The focus embraced an evaluation of different phases of the statistical process. First, a methodological work to explore the pre-processing step of DNA methylation data was implemented. Its aim was to first identify the various sources of systematic and random variability, related to sample treatment, laboratory, as well as biological, and then to screen among the most popular normalization techniques. The PC-PR2 method proved a useful tool to explore the contribution of an *a priori* list of factors in large dimension datasets, such as epigenetics data. For the normalization phase, the SVA technique produced more conservative results than the two other methods investigated, i.e. Combat and a method based on residuals computation, possibly in light of the fact that SVA makes use of the notion of surrogate variables, thus correcting for what is known to affect variation, but also involving unknown sources of variability.

Three statistical methods were described in this thesis to analyze methylation data in order to investigate the association between dietary folate, alcohol intake and DNA methylation. The site-specific analysis, where single CpG sites were independently related to, in turn, alcohol and folate, served as a basis to go beyond ‘univariate’ evaluations of the relationships. The DMRs and FL analyses provided evidence that specific regions of CpG sites were associated with lifestyle factors using the hypothesis that neighboring features may share similar information. DMRs and FL analyses indicated that dietary folate and alcohol intake might be associated with alteration of DNA methylation in localized regions, some of which are related to genes known to act as tumor suppressor. These results were in line with the hypothesis that epigenetic mechanisms might have a role in the association between folate and alcohol with BC. A fourth study investigating the relationship between

global methylation and the risk of BC was also presented, and showed that overall global methylation was not associated with BC risk, whereas a positive association was observed in CpG islands. It is important to stress that the associations observed in this thesis should be interpreted with caution, as our findings need confirmation in other study populations in similar research settings.

Perspectives

Many environmental exposure including smoking, obesity and specific dietary factors are suspected to contribute to methylation changes, which may entail the development of a range of chronic diseases such as cardiovascular disease, type-I diabetes and several cancer types, including colorectal and lung. By addressing the high dimensionality and complexity of DNA methylation, statistical tools introduced in this thesis may prove useful for future epigenetics studies focusing on the relationship between lifestyle exposures, DNA methylation and the occurrence of health outcomes.

Among fatty acids profiles, positive associations have been recently observed between plasma palmitoleic acid, as a biomarker of endogenous lipogenesis, and BC risk, and also between industrial trans-fatty acids and ER-negative breast tumours (62). Fatty acids are suspected to alter BC risk through an hypo-methylation of specific CpG sites, possibly resulting from an alteration of the activities of the TET proteins and a reduced DNA methyltransferases activity (90).

A study aiming at investigating the association between biomarkers of endogenous lipogenesis, DNA methylation and BC is currently ongoing. The rationale of this investigation is to use DMR analysis to identify CpG regions showing altered methylation levels altered by specific fatty acid biomarkers. In addition, the association between methylation levels and BC risk will be assessed in each CpG region, by summarizing methylation intensity of the CpG sites belonging to the region by reduction dimension techniques, and then relating the resulting components (or factors) to the risk of BC. Analyses for palmitoleic acid were performed, while analyses for industrial trans-fatty acids are currently ongoing. Plasma palmitoleic acid was associated with methylation changes in 48 DMRs (annex 1). Methylation levels from CpG sites in 11 DMRs were significantly associated with BC risk (annex 2).

Statistical tools presented in this thesis may also be extended to other types of *-omics* data. Some of the statistical methods may need to be adapted to suite the specific setting of large dimension data. For example, as some of the *-omics* data are not ordered, analyses involving the concept of physical proximity of features, such DMR and FL regression, may

not find a straightforward application. A potential extension of DMR analysis may be adapted to identify cluster of features associated with an exposure, by using weights based on correlation between features instead of weights based on physical distance between CpG sites. For FL regression, a pre-step would be needed to order features, possibly using hierarchical clustering methods.

REFERENCES

1. Chadeau-Hyam M*, et al.* (2013) Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. *Environmental and molecular mutagenesis* 54(7):542-557.
2. Vineis P, Veldhoven K, Chadeau-Hyam M, & Athersuch Toby J (2013) Advancing the application of omics-based biomarkers in environmental epidemiology. *Environmental and molecular mutagenesis* 54(7):461-467.
3. Fages A*, et al.* (2014) Investigating sources of variability in metabolomic data in the EPIC study: the Principal Component Partial R-square (PC-PR2) method. *Metabolomics* 10(6):1074-1083.
4. Leek JT (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews. Genetics* 11.
5. Harper KN, Peters BA, & Gamble MV (2013) Batch effects and pathway analysis: two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiol Biomarkers Prev* 22(6):1052-1060.
6. Jiao C*, et al.* (2018) Positional effects revealed in Illumina methylation array and the impact on analysis. *Epigenomics* 10(5):643-659.
7. Verdugo RA, Deschepper CF, Muñoz G, Pomp D, & Churchill GA (2009) Importance of randomization in microarray experimental designs with Illumina platforms. *Nucleic Acids Research* 37(17):5610-5618.
8. Kitchen RR*, et al.* (2011) Relative impact of key sources of systematic noise in Affymetrix and Illumina gene-expression microarray experiments. *BMC Genomics* 12:589.
9. Johnson WE, Li C, & Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics (Oxford, England)* 8.
10. Sims AH*, et al.* (2008) The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Medical Genomics* 1(1):1-14.
11. Benito M*, et al.* (2004) Adjustment of systematic microarray data biases. *Bioinformatics (Oxford, England)* 20(1):105-114.
12. Leek JT & Storey JD (2007) Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet* 3(9):e161.
13. Leek JT & Storey JD (2008) A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences of the United States of America* 105(48):18718-18723.
14. Hoerl AE & Kennard RW (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* 12(1):55-67.
15. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Roy Stat Soc B* 58.
16. Zou H & Hastie T (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301-320.
17. Jolliffe IT (1982) A Note on the Use of Principal Components in Regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31(3):300-303.
18. Tenenhaus M (1998) *La régression PLS: théorie et pratique* (Editions Technip).
19. Abdi H (2010) Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdisciplinary Reviews: Computational Statistics* 2(1):97-106.
20. Peters TJ*, et al.* (2015) De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin* 8(1):1-16.
21. Waddington CH (1942) CANALIZATION OF DEVELOPMENT AND THE INHERITANCE OF ACQUIRED CHARACTERS. *Nature* 150:563.

22. Felsenfeld G (2014) A Brief History of Epigenetics. *Cold Spring Harbor Perspectives in Biology* 6(1):a018200.
23. Berger SL, Kouzarides T, Shiekhattar R, & Shilatifard A (2009) An operational definition of epigenetics. *Genes & development* 23(7):781-783.
24. Heard E, Chaumeil J, Masui O, & Okamoto I (2004) Mammalian X-chromosome inactivation: an epigenetics paradigm. *Cold Spring Harbor symposia on quantitative biology* 69:89-102.
25. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature reviews. Genetics* 13(7):484-492.
26. Moore LD, Le T, & Fan G (2013) DNA Methylation and Its Basic Function. *Neuropsychopharmacology* 38(1):23-38.
27. Aran D, Toperoff G, Rosenberg M, & Hellman A (2011) Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet* 20.
28. Bibikova M*, et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics* 98(4):288-295.
29. Ehrlich M*, et al.* (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res* 10(8):2709-2721.
30. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes & development* 16(1):6-21.
31. Zaidi SK*, et al.* (2010) Architectural epigenetics: mitotic retention of mammalian transcriptional regulatory information. *Molecular and cellular biology* 30(20):4758-4766.
32. Lillycrop KA & Burdge GC (2015) Maternal diet as a modifier of offspring epigenetics. *Journal of developmental origins of health and disease* 6(2):88-95.
33. Feil R & Fraga MF (2011) Epigenetics and the environment: emerging patterns and implications. *Nature reviews. Genetics* 13(2):97-109.
34. Heyn H*, et al.* (2012) Distinct DNA methylomes of newborns and centenarians. *Proc Natl Acad Sci USA* 109.
35. Ambatipudi S*, et al.* (2016) Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* 8(5):599-618.
36. Joehanes R*, et al.* (2016) Epigenetic Signatures of Cigarette Smoking. *Circulation. Cardiovascular genetics* 9(5):436-447.
37. Wahl S*, et al.* (2017) Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 541(7635):81-86.
38. Wilson LE, Harlid S, Xu Z, Sandler DP, & Taylor JA (2017) An epigenome-wide study of body mass index and DNA methylation in blood using participants from the Sister Study cohort. *International journal of obesity (2005)* 41(1):194-199.
39. Niculescu MD & Haggarty P (2011) *Nutrition in Epigenetics* (Wiley).
40. Burdge G & Lillycrop K (2017) *Nutrition, Epigenetics And Health* (World Scientific Publishing Company, New Jersey).
41. Neidhart M (2015) *DNA Methylation and Complex Human Disease* (Academic Press) p 552.
42. Dang MN, Buzzetti R, & Pozzilli P (2013) Epigenetics in autoimmune diseases with focus on type 1 diabetes. *Diabetes/metabolism research and reviews* 29(1):8-18.
43. Stenvinkel P*, et al.* (2007) Impact of inflammation on epigenetic DNA methylation - a novel risk factor for cardiovascular disease? *Journal of internal medicine* 261(5):488-499.
44. Raftopoulos L*, et al.* (2015) Epigenetics, the missing link in hypertension. *Life sciences* 129:22-26.
45. Kabesch M & Adcock IM (2012) Epigenetics in asthma and COPD. *Biochimie* 94(11):2231-2241.
46. Barrow TM & Michels KB (2014) Epigenetic epidemiology of cancer. *Biochemical and biophysical research communications*.
47. Esteller M (2008) Epigenetics in cancer. *The New England journal of medicine* 358(11):1148-1159.

48. Johansson A & Flanagan JM (2017) Epigenome-wide association studies for breast cancer risk and risk factors. *Trends in cancer research* 12:19-28.
49. Coppede F (2014) Epigenetic biomarkers of colorectal cancer: Focus on DNA methylation. *Cancer letters* 342(2):238-247.
50. Liloglou T, Bediaga NG, Brown BR, Field JK, & Davies MP (2014) Epigenetic biomarkers in lung cancer. *Cancer letters* 342(2):200-212.
51. Galm O, Herman JG, & Baylin SB (2006) The fundamental role of epigenetics in hematopoietic malignancies. *Blood reviews* 20(1):1-13.
52. Ferlay J*, et al.* (2014) Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer. Journal international du cancer*.
53. Allemani C*, et al.* (2015) Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet* 385(9972):977-1010.
54. Tretli S (1989) Height and weight in relation to breast cancer morbidity and mortality. A prospective study of 570,000 women in Norway. *International journal of cancer. Journal international du cancer* 44(1):23-30.
55. Pharoah PD, Day NE, Duffy S, Easton DF, & Ponder BA (1997) Family history and the risk of breast cancer: a systematic review and meta-analysis. *International journal of cancer. Journal international du cancer* 71(5):800-809.
56. Singletary SE (2003) Rating the risk factors for breast cancer. *Annals of surgery* 237(4):474-482.
57. Romieu I*, et al.* (2015) Alcohol intake and breast cancer in the European prospective investigation into cancer and nutrition. *International journal of cancer. Journal international du cancer* 137(8):1921-1930.
58. Wu Y, Zhang D, & Kang S (2013) Physical activity and risk of breast cancer: a meta-analysis of prospective studies. *Breast cancer research and treatment* 137(3):869-882.
59. World Cancer Research Fund International & American Institue for Cancer Research (2017) Continuous Update Project Report: Diet, nutrition, physical activity and breast cancer.
60. Chajes V & Romieu I (2014) Nutrition and breast cancer. *Maturitas* 77(1):7-11.
61. de Batlle J*, et al.* (2015) Dietary folate intake and breast cancer risk: European prospective investigation into cancer and nutrition. *Journal of the National Cancer Institute* 107(1):367.
62. Chajes V*, et al.* (2017) A prospective evaluation of plasma phospholipid fatty acids and breast cancer risk in the EPIC study. *Annals of oncology : official journal of the European Society for Medical Oncology*.
63. Teegarden D, Romieu I, & Lelievre SA (2012) Redefining the impact of nutrition on breast cancer incidence: is epigenetics involved? *Nutrition research reviews* 25(1):68-95.
64. Michels KB, Mohllajee AP, Roset-Bahmanyar E, Beehler GP, & Moysich KB (2007) Diet and breast cancer: a review of the prospective observational studies. *Cancer* 109(12 Suppl):2712-2749.
65. Szyf M (2011) The implications of DNA methylation for toxicology: toward toxicomethylomics, the toxicology of DNA methylation. *Toxicological sciences : an official journal of the Society of Toxicology* 120(2):235-255.
66. Rampersaud GC, Kauwell GP, Hutson AD, Cerda JJ, & Bailey LB (2000) Genomic DNA methylation decreases in response to moderate folate depletion in elderly women. *The American journal of clinical nutrition* 72(4):998-1003.
67. Kruman, II & Fowler AK (2014) Impaired one carbon metabolism and DNA methylation in alcohol toxicity. *Journal of neurochemistry* 129(5):770-780.
68. Chen J*, et al.* (2005) One-carbon metabolism, MTHFR polymorphisms, and risk of breast cancer. *Cancer research* 65(4):1606-1614.

69. Zhang S*, et al.* (1999) A prospective study of folate intake and the risk of breast cancer. *Jama* 281(17):1632-1637.
70. Zhang SM*, et al.* (2003) Plasma folate, vitamin B6, vitamin B12, homocysteine, and risk of breast cancer. *Journal of the National Cancer Institute* 95(5):373-380.
71. Liu C*, et al.* (2016) A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry*.
72. Mason JB & Choi S-W (2005) Effects of alcohol on folate metabolism: implications for carcinogenesis. *Alcohol* 35(3):235-241.
73. Stolzenberg-Solomon RZ*, et al.* (2006) Folate intake, alcohol use, and postmenopausal breast cancer risk in the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial. *The American journal of clinical nutrition* 83(4):895-904.
74. Beasley JM*, et al.* (2010) Alcohol and risk of breast cancer in Mexican women. *Cancer causes & control : CCC* 21(6):863-870.
75. Riboli E*, et al.* (2002) European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr* 5(6B):1113-1124.
76. Matejcic M*, et al.* (2017) Biomarkers of folate and vitamin B12 and breast cancer risk: report from the EPIC cohort. *International journal of cancer. Journal international du cancer* 140(6):1246-1259.
77. Ambatipudi S*, et al.* (2017) DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. *European Journal of Cancer* 75:299-307.
78. Herceg Z (2016) Epigenetic Mechanisms as an Interface Between the Environment and Genome. *Advances in experimental medicine and biology* 903:3-15.
79. Ames BN & Wakimoto P (2002) Are vitamin and mineral deficiencies a major cancer risk? *Nature reviews. Cancer* 2(9):694-704.
80. Vera-Ramirez L*, et al.* (2013) Impact of diet on breast cancer risk: a review of experimental and observational studies. *Critical reviews in food science and nutrition* 53(1):49-75.
81. Lin J*, et al.* (2008) Plasma folate, vitamin B-6, vitamin B-12, and risk of breast cancer in women. *The American journal of clinical nutrition* 87(3):734-743.
82. Wu K*, et al.* (1999) A prospective study on folate, B12, and pyridoxal 5'-phosphate (B6) and breast cancer. *Cancer Epidemiol Biomarkers Prev* 8(3):209-217.
83. Ba Y*, et al.* (2011) Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II in maternal and cord blood. *European journal of clinical nutrition* 65(4):480-485.
84. Tibshirani R, Saunders M, Rosset S, Zhu J, & Knight K (2005) Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1):91-108.
85. Benz CC (2008) Impact of aging on the biology of breast cancer. *Critical reviews in oncology/hematology* 66(1):65-74.
86. Horvath S (2013) DNA methylation age of human tissues and cell types. *Genome Biol* 14(10):R115.
87. Levine ME, Lu AT, Bennett DA, & Horvath S (2015) Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging* 7(12):1198-1211.
88. Horvath S & Ritz BR (2015) Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging* 7(12):1130-1142.
89. Levine ME*, et al.* (2015) DNA methylation age of blood predicts future onset of lung cancer in the women's health initiative. *Aging* 7(9):690-700.
90. Burdge GC & Lillycrop KA (2014) Fatty acids and epigenetics. *Current opinion in clinical nutrition and metabolic care* 17(2):156-161.

ANNEXES

Annex 1. The 15 of the most significant DMRs associated with palmitoleic acid out of 48 significant DMRs.

#	DMRs characteristics				CpGs characteristics[1]	
	Associated genes	Gene regions	hg19coord	Sites[2]	q_{DMR}[3]	β_{DMR}[4]
1	PM20D1	Body, 1stExon, 5'UTR, TSS200, TSS1500	chr1:205818484-205819609	9	5,1E-10	-0,240
2	HLA-F	Body	chr6:29691643-29692995	20	8,4E-06	-0,072
3	CAT	TSS1500, TSS200, 1stExon, 5'UTR, Body	chr11:34460107-34461028	10	3,1E-04	-0,140
4			chr8:144437314-144437914	6	4,2E-04	0,102
5	NAPRT1	Body,1stExon, TSS200, TSS1500	chr8:144659831-144661051	7	7,2E-04	0,157
6	TNXB	Body	chr6:32054561-32055738	27	7,4E-04	-0,044
7	CD8A	TSS1500	chr2:87036626-87037038	4	1,2E-03	0,089
8	RASA3	Body	chr13:114800796-114801587	6	1,3E-03	-0,120
9	L3MBTL	TSS1500, TSS200, 5'UTR, 1stExon	chr20:42142224-42143211	24	1,3E-03	-0,031
10	FAM196A, DOCK1	1stExon, 5'UTR, Body, TSS200, TSS1500	chr10:128994297-128995192	9	1,3E-03	0,066
11	SHANK2	1stExon, Body, TSS200, TSS1500	chr11:70507825-70508659	11	1,4E-03	-0,080
12	LIME1	Body	chr20:62368956-62369605	5	1,4E-03	0,086
13	C16orf11	TSS1500	chr16:609353-609679	2	1,4E-03	-0,044
14	PCDHGA cluster [5]	Body,5'UTR,1stExon	chr5:140810260-140811102	8	1,6E-03	-0,082
15	GPR75, LOC100302652	5'UTR, Body, 1stExon, TSS200, TSS1500	chr2:54086854-54087552	14	1,8E-03	0,113

[1] Adjusted for alcohol intake, recruitment centre, age at recruitment, menopausal status and level of different lymphocyte subtypes;
[2] Number of sites located in DMRs significant for palmitoleic acid;
[3] Minimum palmitoleic acid q-values of sites located in the DMRs (FDR correction);
[4] Absolute maximum of palmitoleic acid coefficient of sites located in the DMRs;
[5] PCDHGA cluster of genes including : PCDHGA4, PCDHGA11, PCDHGA12, PCDHGA9, PCDHGA1, PCDHGB1, PCDHGB6, PCDHGB3, PCDHGB7, PCDHGA6, PCDHGA8, PCDHGA10, PCDHGA5, PCDHGB4, PCDHGA3, PCDHGA2, PCDHGA7, PCDHGB2, PCDHGB5.

Annex 2. DMRs significantly associated with BC risk.

DMR characteristics			Model 1[1]	Model 2[1]	
#	Associated genes	Sites[2]	Sites sign[3]	# PC[4]	PC sign[5]
1	PM20D1	9	0	1	PC1
3	CAT	10	1	3	0
6	TNXB	27	4	12	PC6
9	L3MBTL	24	1	12	PC7
14	PCDHGA cluster [6]	8	1	3	PC2
15	GPR75,LOC100302652	14	0	3	PC1
16		7	1	3	0
19		6	4	4	PC3, PC4
20	PPT2	17	3	6	PC6
22	WSCD1	3	1	2	0
30	FAM171A2	5	2	1	0
34	FAM38A	6	2	4	PC4
35	ZNF232	3	1	2	0

[1] BC risk was regressed on methylation levels of all the CpG sites included in the DMRs for model 1 and on PC scores keeping 80% of information for model 2. Adjustment covariates were alcohol intake, BMI and physical activity;
[2] Number of sites located in DMRs significant for palmitoleic acid;
[3] Number of CpG sites significantly associate with BC risk;
[4] Number of principal components (PC) needed to keep 80% of information in PCA;
[5] Principal components significantly associate with BC risk using model 2;
[6] PCDHGA cluster of genes including : PCDHGA4, PCDHGA11, PCDHGA12, PCDHGA9, PCDHGA1, PCDHGB1, PCDHGB6, PCDHGB3, PCDHGB7, PCDHGA6, PCDHGA8, PCDHGA10, PCDHGA5, PCDHGB4, PCDHGA3, PCDHGA2, PCDHGA7, PCDHGB2, PCDHGB5.