Université de Montréal

DNA methylation of *F2RL3* and *AHRR* and lung cancer risk

Par

Alice Nguyen

Département de médecine sociale et préventive

École de santé publique

Mémoire présenté en vue de l'obtention du grade de Maîtrise

En Épidémiologie

Décembre 2022

Université de Montréal

Ce mémoire intitulé

DNA methylation of *F2RL3* and *AHRR* and lung cancer risk

Présenté par

Alice Nguyen

A été évaluée par un jury composé des personnes suivantes :

Isabelle Doré

Président-rapporteur

Vikki Ho

Directrice de recherche

Anita Koushik

Codirectrice

Helen Trottier

Membre du jury

Résumé

Introduction: L'étude des biomarqueurs a le potentiel de documenter sur les mécanismes sous-jacents de l'étiologie du cancer du poumon. Dans cette étude, nous avons étudié l'association entre la méthylation de l'ADN dans les gènes *F2RL3* et AHRR et le cancer du poumon.

Méthodes: Une étude cas-témoin avec échantillonnage cumulatif a été nichée dans la cohorte CARTaGENE. Les cas (N=187) se composent de tous les participants diagnostiqués avec un cancer du poumon incident entre le début de la cohorte (2009) et 2015 et qui avaient fourni un échantillon de sang; les témoins (N=378) ont été échantillonnés à la fin du suivi parmi les non-malades selon un appariement fréquentiel (2 :1) pour l'âge, le sexe et le moment du prélèvement sanguin. Sequenom EpiTYPER® a été utilisé pour quantifier les niveaux de méthylation dans sept et 33 sites CpG de *F2RL3* et *AHRR*, respectivement. Les rapports de méthylation de l'ADN sur tous les sites CpG individuels et en tant que mesure moyenne ont été paramétrés à la fois comme variables continues et catégorielles. Une régression logistique multivariable non conditionnelle a été utilisée pour estimer les rapports de cotes (OR) et les intervalles de confiance (IC) à 95 % de l'association entre la méthylation de *F2RL3* et *AHRR* et le cancer du poumon tout en contrôlant les facteurs de confusion identifiés à l'aide de graphiques acycliques dirigés.

Résultats: Une forte association inverse entre les niveaux moyens de méthylation de l'ADN et le cancer du poumon a été observée pour *F2RL3* (OR par écart type (SD) de changement de méthylation = 0,65, IC à 95 % : 0,53-0,80) et *AHRR* (OR par SD de changement de méthylation = 0,66, IC à 95 % : 0,53 à 0,80). De même, les sites CpG individuels ont montré des ORs (par SD de changement de méthylation) allant de 0,61 à 0,70 pour six des sept sites CpG de *F2RL3* et de 0,57 à 0,79 pour 17 des 33 sites CpG de *AHRR*. Les sites CpG restants de *F2RL3* et *AHRR* n'ont

montré aucune association avec le risque de cancer du poumon, à l'exception d'un site CpG dans *AHRR* (chr5:369774) qui avait un OR de 1,25 (IC à 95 % : 1,02-1,54).

Conclusion : Ces résultats confirment le rôle des mécanismes épigénétiques dans l'étiologie du cancer du poumon.

Mots-clés : Cancer du poumon, épigénétique, méthylation de l'ADN, répresseur du récepteur d'aryl hydrocarbone, facteur de coagulation II (thrombine) receptor-like 3

Abstract

Background: The study of biomarkers has the potential to inform on underlying mechanisms in lung cancer etiology. In this study, we investigated DNA methylation in the *F2RL3* and *AHRR* genes, and lung cancer risk.

Methods: A case-control study with cumulative sampling was nested in the CARTaGENE cohort. Cases (N=187) consisted of all participants diagnosed with incident lung cancer from baseline to 2015 and who had provided a blood sample; controls (N=378) were sampled at a ratio of 2:1 with frequency-matching by age, sex, and timing of blood sampling. Sequenom EpiTYPER® was used to quantify methylation levels in seven and 33 CpG sites of *F2RL3* and *AHRR,* respectively. DNA methylation ratios across all individual CpG sites and as an average measure were parametrized both as continuous and categorical variables. Unconditional multivariable logistic regression was used to estimate odds ratios (ORs) and 95% confidence intervals (CI) for lung cancer associated with *F2RL3* and *AHRR* methylation while controlling for confounders identified using directed acyclic graphs.

Results: A strong inverse relationship between average DNA methylation levels and lung cancer was observed for both *F2RL3* (OR per standard deviation (s.d.) in methylation change = 0.65, 95% CI: 0.53-0.80) and *AHRR* (OR per s.d. in methylation change = 0.66, 95% CI: 0.53-0.80). Similarly, ORs for individual CpG sites (per s.d. in methylation change) ranged from 0.61-0.70 for six out of the seven CpG sites of *F2RL3* and from 0.57-0.79 for 17 out of 33 CpG sites of *AHRR.* The methylation levels of the remaining CpG sites within *F2RL3* and *AHRR* were not associated with lung cancer risk, except for one CpG site within *AHRR* (chr5:369774) which had an OR of 1.25 (95% CI: 1.02-1.54).

Conclusion: These findings support the role of epigenetic mechanisms in lung cancer etiology.

Keywords: Lung cancer, epigenetics, DNA methylation, aryl-hydrocarbon receptor repressor, coagulation factor II (thrombin) receptor-like 3

Table of contents

List of tables

List of figures

List of acronyms

AhR: Ah Receptor pathway

AHR: Aryl Hydrocarbon Receptor

AHRR: Aryl Hydrocarbon Receptor Repressor

ARNT: Aryl Hydrocarbon Nuclear Translocator

BH: Benjamini-Hochberg

BMI: Body Mass Index

CDC: Centers for Disease Control and Prevention

CER: Comité d'Éthique de la Recherche

CHUM: Centre Hospitalier de l'Université de Montréal

CI: Confidence interval

CIHR: Canadian Institute of Health Research

CpG: Cytosine-phosphate-Guanine

CSI: Cumulative Smoking Index

CV: Coefficient of Variation

DAG: Directed Acyclic Graph

DNA: Deoxyribonucleic acid

EWAS: Epigenome-Wide Association Study

HR: Hazard Ratio

LDCT: Low-Dose Computed Tomography

MAR: Missing at Random

OR: Odd Ratio

PAR-4: Protease-Activated Receptor-4

PC: Principal Component

PCA: Principal Component Analysis

PCR: Polymerase Chain Reaction

RAMQ: Régie de l'Assurance Maladie du Québec

RNA: Ribonucleic acid

RR: Relative risk

SES: Socioeconomic Status

SNP: Single Nucleotide Polymorphism

List of abbreviations

e.g.: From the Latin, Exempli gratia (For example)

et al.: From the Latin, et alia (and others.)

i.e.: From the Latin, Id est (That is)

s.d.: standard deviation

Acknowledgements

I would like to express my sincerest gratitude to my supervisor, Dr Vikki Ho, who made this project possible. Her guidance and patience during challenging times has given me the will and confidence to finish this project. I would also like to extend my thanks to my co-supervisor, Dr Anita Koushik, for sharing her expertise and providing insightful knowledge.

I am thankful for my fellow labmates in Dr. Ho's and Dr. Koushik's lab for their companionship throughout this journey. Especially, Michael, Laura, and Chelsea who I could always count on for moral support. I am also grateful for my friends in the Public Health Master's program, Elizabeth, Sarah, and Simon, who have made this process enjoyable.

This journey would not have been possible without the constant support and love from my family, especially my dad, mom, and sister. Last but not least, thank you to my partner, Fred, for believing in me through all the ups and downs.

In the loving memory of my grandparents,
Thị Huyền Lê and Tiến Lê,
who both bravely fought against cancer

Chapter 1. Introduction

Despite progress in lung cancer treatment, both primary and secondary prevention continues to be important for this disease with a high mortality rate and considerable burden on the patient. Although smoking is the principal risk factor, only a proportion of smokers develop lung cancer and approximately 10-15% of incident cases occur in never-smokers (1). A better understanding of risk factors and underlying causal mechanisms is thus necessary to support prevention initiatives. The study of intermediate endpoints or events has enormous potential with respect to understanding lung cancer etiology. The relationship between intermediate endpoints and lung cancer must thus be established in order to effectively use such markers in population health studies.

Accumulating evidence supports the role of DNA methylation in cancer development; hence, its use as an intermediate carcinogenic event in lung cancer development shows promise (2). DNA methylation is an epigenetic process whereby a methyl group is transferred to a cytosine residue in sites where it is followed by a guanine residue, also called CpG sites (3). Global genome-wide methylation and specific hyper- or hypomethylation of CpG rich regions (or CpG islands) in promoters of particular genes are two forms of aberrant DNA methylation found in lung cancer. The former has been mostly associated with cancer progression, a late event in cancer development while the latter has been linked to early lung cancer initiation (2). The exploration of gene-specific methylation in relation to lung cancer could therefore be of relevance to identify potential early markers of lung carcinogenesis.

In an epigenome-wide association study (EWAS) of four pooled prospective cohorts, hypomethylation of specific CpG sites in the smoking-related genes *F2RL3* and *AHRR* were found to be strongly associated with increased lung cancer risk (4). *F2RL3*, or the coagulation factor II

receptor like 3 gene, encodes for the protease-activated receptor-4 (PAR-4) known to influence blood coagulation and immune response and to be involved in neoplastic diseases (5,6). *AHRR* encodes for the aryl hydrocarbon receptor repressor (AHRR) and is involved in the inflammatory response, apoptosis, and cell proliferation (7,8). This thesis thus, sought to determine whether methylation levels in additional CpG sites of these two genes are associated with lung cancer risk in order to inform on the role of methylation as an intermediate endpoint in lung cancer etiology.

This thesis is organized into seven chapters. An overview of lung cancer burden and the current state of knowledge in relation to the role of DNA methylation in lung cancer etiology are introduced in Chapter 2. Chapter 3 presents this study's objective and hypothesis. Chapter 4 describes the methodology used in this thesis. Chapter 5 is the manuscript that will be submitted to the "Cancer Epidemiology, Biomarkers & Prevention" journal. Supplementary results are presented in Chapter 6. Summaries of our study findings and the strengths and limitations of the study, as well as future directions, are discussed in Chapter 7.

Chapter 2. Literature review

2.1 Burden of lung cancer

The global proportion of deaths attributed to non-communicable diseases is predicted to increase from 59% in 2002 to 69% by 2030 (9). Cancer is projected to account for a quarter of those deaths, making it the second leading cause of death after cardiovascular diseases (9). Furthermore, lung cancer is currently the second most common cancer in both men and women globally with an estimated 2.21 million incident cases in 2020 (10,11). It is also the leading cause of cancer death, accounting for the 18% of the global cancer-related mortality burden in 2020 (10). In Canada, 29,600 Canadiens are estimated to have been diagnosed with lung cancer in 2021, representing up to 13% of incident cancer cases (12). One in 19 Canadians are estimated to die from lung cancer, as lung cancer also remains the leading cause of cancer death in Canada, representing approximately 25% of all cancer deaths (13).

Lung cancer incidence has been on a downward trend, closely following the decrease in smoking after the introduction of cessation programs globally (14). Nevertheless, the global lung cancer mortality rate remains dismally high (14), with more than half of lung cancer cases dying within one year of being diagnosed (15). This poor survival rate is largely determined by the late stages (stage III/IV) at which 75% of lung cancer cases are diagnosed (16). Lung cancers detected at an early and localized stage (stage I) have a five-year survival rate of 52% compared to 24.3% and 3.6% for lung cancers diagnosed at later stages where the cancer has spread to regional or distant parts of the body (stage II/III and IV, respectively) (15). The low survival rates along with the late stages at which lung cancer cases are diagnosed highlights the need for continued research on alternative screening methods to effectively support secondary prevention initiatives.

2.2 Overview of lung cancer risk factors

While tobacco smoking is the leading risk factor for lung cancer, additional risk factors are implicated in lung carcinogenesis. The following sections present an overview of established lung cancer risk factors.

2.2.1 Tobacco smoking

Tobacco smoking is one of the most established risk factors for lung cancer. The World Health Organisation has estimated that the global smoking pattern of 5.6 trillion cigarettes smoked per year will be the cause of more than 8 million lung cancer deaths by 2030 (17). A pooled analysis of 13,169 cases and 16,010 controls from seven case-control studies in Europe and Canada showed that current smokers have higher odds of having lung cancer compared to never smokers (men: OR: 23.6, 95% CI: 20.4-27.2; women: OR: 7.8, 95% CI: 6.8-9.0) (18). In the same study, it was found that the odds of lung cancer among former smokers never decreases down to those observed for never smokers even 35 years after smoking cessation (men: OR: 7.5, 95% CI: 6.5-8.7; women: OR: 2.8, 95% CI: 2.4-3.3) (18). Results from an analysis within the Framingham cohort study showed that while lung cancer risk drops for heavy former smokers within five years since quitting relative to current smokers (Hazards ratio [HR]: 0.61, 95% CI: 0.40-0.93), this risk remains more than threefold higher than never smokers even 25 years after quitting (HR: 3.85, 95% CI: 1.80-8.26) (19).

2.2.2 Age

Lung cancer is most often diagnosed in patients over 65 years old, accounting for approximately 69% of incident cases from 2004 to 2008 (11). In Canada, the incidence rate of lung cancer peaks among the population aged 75 to 84 (13).

2.2.3 Sex

Several studies have shown that women are at a higher risk of lung cancer than men (1,20). Specifically, a review that presented new data on never-smokers from six large population-based cohorts has described that age-adjusted incidence rates of lung cancer for women (14.4 to 20.8 per 100,000 person-years) were higher than men (4.8 to 13.7 per 100,000 person-years) (1). The disparity in lung cancer risk by sex among never-smokers has been suggested to be due to differences by sex in molecular (such as sex hormones and genetic factors) and environmental exposures (such as second-hand smoke) (21,22).

2.2.4 Fruits and vegetable consumption

Dietary factors are important risk factors for lung cancer. A high intake of fruits and vegetables has been found to be inversely associated with lung cancer risk. A meta-analysis of 13 cohort studies and 19 case-control studies reported a pooled relative risk (RR) for lung cancer of 0.74 (95% CI: 0.67-0.82) when comparing the highest versus lowest intake of fruits and vegetables (23). Stratified by study design, RRs of 0.88 (95% CI: 0.81-0.97) and 0.62 (95% CI: 0.54-0.70) were estimated for cohort and case-control studies, respectively, when comparing the highest versus lowest intake categories of vegetable intake (23). In the same study, the pooled RR of lung cancer for the highest category of fruit intake versus the lowest category was 0.84 (95% CI: 0.75-0.94) and 0.77 (95% CI: 0.67-0.88) for cohort and case-control studies, respectively (23). A similar reduction in lung cancer risk was observed in a meta-analysis that looked at the association of fruit and vegetable intake and lung cancer risk among participants with different smoking statuses (24).

2.2.5 Body mass index

Contrary to other types of cancer, obesity has been unexpectedly observed to be inversely associated with lung cancer risk, and this surprising observation is often referred to as the *obesity*

paradox (25–28). After stratification by study design, a meta-analysis of 20 cohort studies and 11 case-control studies reported an inverse association between BMI and lung cancer risk. Participants in the overweight (body mass index, BMI = 25-29.925 kg/m^2) and obese (BMI $\geq$ 25 kg/m^2) categories had a lower risk for lung cancer relative to participants in the normal weight category (BMI = 18.5-24.9 kg/m^2) for both cohorts studies (overweight: RR = 0.78, 95% CI: 0.72-0.84; obese: RR = 0.80, 95% CI: 0.73-0.88) and case-control studies (overweight: RR = 0.68, 95% CI: 0.57-0.82; obese: RR = 0.56, 95% CI: 0.40-0.79) (27). An additional BMI category named as "excess body weight" (BMI $\geq$ 25 kg/m^2), which combines both the overweight and obese categories, was considered by the same meta-analysis. Similarly to the overweight and obese categories, participants with a BMI considered as "excess body weight" had a lower risk for lung cancer relative to participant in the "normal" BMI category for both cohorts studies (RR = 0.78, 95% CI: 0.72-0.84) and case-control studies (RR = 0.65, 95% CI: 0.52-0.81) (27). Interestingly, following stratification by smoking status, this relationship was found to be attenuated in non-smokers but strengthened in current and former smokers (26–28). A prospective study that has looked directly at the question of whether the inverse association between BMI and lung cancer risk is due to residual confounding by smoking within previous studies has found no evidence to support that theory (28).

2.2.6 Ethnicity

There is growing evidence for ethnic differences in lung cancer incidence (29). In the United States, a report by the Centers for Disease Control and Prevention (CDC) looking at data from 1998 to 2006 from the Surveillance, Epidemiology, and End Results Program and from the CDC's National Program of Cancer Registries revealed that Blacks have a higher lung cancer incidence than other ethnic groups (30). Specifically, they reported an annual incidence per

100,000 of 76.1 in Blacks, 69.7 in Whites, 48.4 in American Indians/Alaska Natives, 38.4 in Asian/Pacific Islanders, and 37.3 in Hispanics (30). While this ethnic disparity in lung cancer was not found to be explained by difference in genetic mutations (31), it has been suggested that this disparity could instead be due to differences in gene expression (32).

2.3 The importance of secondary prevention in reducing the burden of lung cancer

The public health concept of prevention is usually separated into three distinctive categories: primary, secondary, and tertiary. Primary prevention consists of measures that aim to prevent the onset of illness, such as vaccination and smoking cessation programs. Secondary prevention involves the early detection of illness to initiate early treatment and the stalling of its progression. Screening programs such as regularly scheduled mammography exams to detect breast cancer are a good example of a secondary prevention initiative. Tertiary prevention focuses on minimizing the effects of a disease on a person to improve their quality of life (33–35).

Secondary prevention is of interest for lung cancer given the majority of cases are diagnosed at a later stage. Current screening technologies for lung cancer include chest X-ray, sputum cytology, and low-dose computed tomography (LDCT) scanning. A systematic review and meta-analysis of seven randomized controlled trials in the general adult populations has found that current evidence does not support the combined use of chest X-ray and sputum cytological screening as a small 12% risk reduction was observed which was not statistically significant (RR: 0.88, 95%CI: 0.74-1.03) (36). One national randomized controlled trial in the US, which included 53,454 high-risk and former smokers, has demonstrated that the annual LDCT screening group had a lower RR of death compared to the annual chest X-ray screening group after six years of follow-up (RR: 0.80, 95% CI:0.70-0.92) (37). The use of LDCT screening however also comes with its risks. In the same study, LDCT has been shown to have a high rate of false positives

(96.4%) and a rate of 18.50% of overdiagnosis, which is the diagnosis of diseases that would not have caused any symptoms or death if left untreated (37,38). The potential physical and mental harm to an individual and to society from overdiagnosis and a high false positive rate, such as cost and subsequent invasive testing and treatment, must be considered with the continued use of LDCT screening. A better understanding of the underlying mechanisms of lung cancer may inform future development of better screening tests.

2.4 DNA methylation as a biomarker of lung carcinogenesis

The study of biomarkers of intermediate endpoints to facilitate secondary prevention efforts is of accrued interest to further knowledge on the causal mechanism of lung cancer (39) and potentially aid in future screening endeavours. Cancer development and progression are complex processes that are dependent on multiple conditions such as genetic mutations, favorable tumor environment, and epigenetic changes (40). Epigenetic modifications alter gene expression and chromosomal stability without changing the DNA sequence and are reversible, contrary to genetic mutations which are permanent alterations (41). Its disruption has been observed in tandem with genetic changes in cancer, indicating a synergy of both genetic and epigenetic changes to drive the initiation and progression of cancer (42,43). It has also been hypothesized that epigenetic changes are amongst the earliest event in carcinogenesis via the epigenetic disruption of progenitor cells (44).

One epigenetic mechanism highlighted in carcinogenesis is DNA methylation, a process in which a methyl group is transferred to a region of DNA where a cytosine residue is followed by a guanine residue, also termed CpG dinucleotide sites (3,45). Clusters of high CpG density sites near the promoter regions are known as CpG islands. However, regions rich in CpG sites can also be found all over the gene body, and not solely in the promoter region, and those that are 2

kilobases [unit of length of DNA] away from the promoter region are referred to as CpG island shores (46). DNA methylation is essential to normal development due to its role in multiple key biological processes, such as gene expression and regulation, and the disruption of its normal pattern has been commonly observed in human diseases, notably in cancer (41,46). Aberrant DNA methylation is known to be involved in cancer etiology as compelling evidence supports its role in carcinogenesis, and not solely as an event emanating from genetic changes (2,42,44,47).

In cancer, aberrant DNA methylation is characterized as global hypomethylation of the genome and gene-specific hypo- and/or hypermethylation. Global hypomethylation of the genome is more commonly observed in highly repetitive sequences of the DNA, also called DNA repeats. It promotes genomic instability and has been linked to cancer progression, a later event in cancers such as cervical (48), oral (49), and lung (2,50). Specific hyper- or hypo-methylation of CpG sites or CpG islands in or near promoters of certain genes are also types of DNA methylation alterations that have been observed to be involved in carcinogenesis (2). Hypermethylation in the promoter region of a tumor suppressor gene, which drives the silencing of its expression, is often associated with carcinogenesis (51). Similarly, demethylation in or near promoter regions has been linked to cancer through its upregulation of known oncogenes (50).

The study of specific hyper- or hypo-methylation of certain genes, therefore, has potential with regard to lung carcinogenesis. Given that it is measurable in blood (39), DNA methylation is an attractive candidate as a biomarker to aid in secondary prevention efforts.

2.5 Biological mechanism underlying the methylation-lung cancer relationship

Recent studies have reported an association between aberrant methylation of the *F2RL3* and *AHRR* genes in relation to lung cancer risk. *F2RL3*, or the coagulation factor II receptor like 3 gene, encodes for the PAR-4 protein. It is located on the human chromosome 19 and spans 3,608

base pairs of DNA (52). It is thought to be involved in blood coagulation, immune response, and neoplastic diseases (5,6). PAR-4 has also been found to be expressed in several human tissues, including at high levels in normal lung tissues (53). Its function has yet to be elucidated as there is conflicting evidence of its association with lung cancer (4,54,55). PAR-4 is found to be upregulated in cells going through apoptosis, and *in vivo* experiments have demonstrated that loss of PAR-4 instigates the formation of Ras-induced lung carcinoma (56,57). Similarly, a study looking at PAR-4 expression in resected lung adenocarcinoma from 35 patients has reported a decreased expression of PAR-4 compared to adjacent normal lung tissues, supporting the suggestion that PAR-4 acts as a tumor suppressor in lung cancer (54). However, western blotting and immunohistochemical analysis showed that this decrease in PAR-4 expression was more commonly observed amongst later clinical stages of lung cancer (III-IV) (72.1%) compared to earlier stages (I-II) (46.9%) (54). This later observation is in concordance with another study that described an expression of PAR-4 in 39 of the 60 resected stage IB non-small-cell lung cancers, which is considered an early stage. Further research is, thus, still needed to clearly define and clarify the role of PAR-4 in lung cancer.

AHRR encodes for the AHRR protein, a known transcription factor involved in the Ah Receptor (AhR) pathway (8). It is located on the human chromosome 5 and spans 134,116 bps (58). It has been shown to be regulated through DNA methylation, where hypermethylation silences its transcription (8), indicating that hypomethylation of *AHRR* results in its overexpression. AHRR modulates the transcription of AhR-dependent genes by suppressing the activity of the aryl hydrocarbon receptor (AHR) through competition for heterodimer formation with the aryl hydrocarbon nuclear translocator (ARNT) (59,60). AHR is a ligand-induced transcription factor, which is involved in the inflammatory response, apoptosis, and cell

proliferation (7,8). Environmental toxicants such as dioxin and polycyclic aromatic hydrocarbons (PAHs), which are also recognized carcinogens found in tobacco smoke, are known ligands of AHR, demonstrating its role in mediating the toxicity of these xenobiotic compounds through its regulation of their metabolism (61). *AHRR* is recognized as a tumor suppressor gene given its inhibiting effect on AHR transcriptional activity. In other words, hypermethylation of *AHRR*, and thus, the silencing of the expression of the repressor for AHR, permits normal AHR transcriptional activity and metabolism of toxicants, thus preventing an accumulation of toxicants, which could have led to potential genetic damage favorable to cancer development.

Previous studies have examined the relationship between the methylation levels in these two genes with lung cancer risk (Figure 2.1) (4,62). An EWAS of four prospective cohorts by Fasanelli *et al*. has reported inverse associations with lung cancer risk for both the CpG sites cg03636183 in *F2RL3* (OR per s.d. in methylation change: 0.40, 95% CI: 0.31-0.56) and cg05575921 in *AHRR* (OR per s.d. in methylation change: 0.37, 95% CI: 0.31-0.54), indicating hypomethylation of these two genes amongst lung cancer cases as compared to controls (4). This is in agreement with another EWAS by Baglietto *et al*. which looked at five case-control studies nested in prospective cohorts and identified hypomethylation of cg03636183 in *F2RL3* and cg05575921 in *AHRR* in lung cancer cases as compared to controls (62). However, these studies employed epigenome wide methylation assays, such as bead-based microarray technology, to quantitively interrogate the methylation levels of selected CpG sites in the genome. This technology inadvertently only permits these studies to focus on a few CpG sites within the two genes of interest. The study of DNA methylation of multiple CpG sites and their average could potentially further inform on lung cancer risks through a more comprehensive representation of regional methylation levels in both genes, and serve as a better proxy of gene expression (63).

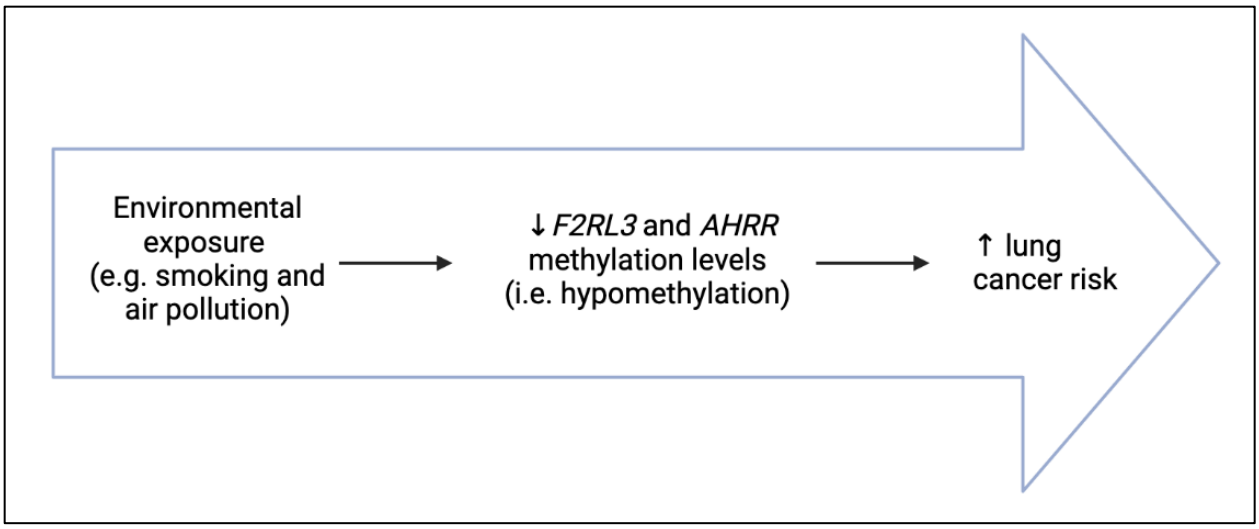

Figure 2.1 Summary of the associations found in relation to *F2RL3* and *AHRR* methylation

2.5.1 Environmental exposure and methylation of *F2RL3* and *AHRR*

There is compelling evidence on the association between smoking, and *F2RL3* and *AHRR* methylation as they were found to be differentially methylated in current smokers compared to never smokers (64,65). An EWAS of 1793 participants has found that current smokers had lower levels of methylation when compared to never-smokers for the CpG sites cg03636183 in *F2RL3* and cg05575921 in *AHRR,* with -14.74% and -24.40% in methylation difference, respectively (65). It was also found that the methylation level of these genes in former smokers gradually increases towards the level of never smokers, proportionally to time since quitting, although former smokers never reach the same methylation levels as those amongst never smokers (66,67).

Other types of exposure have also been associated with lower methylation levels of the same CpG site in the two genes, suggesting that hypomethylation of these genes is not solely related to smoking. Alhamdow, *et al*., has reported that those occupationally exposed to PAHs (i.e. employed as chimney sweeps and among creosote-exposed workers) have lower DNA methylation levels for both *F2RL3* and *AHRR* when compared to unexposed (68). Likewise, a lower methylation level of cg05575921 in *AHRR* was observed in non-smoking Taiwanese adults living

in areas with higher air pollution of fine inhalable particles (also referred to as particulate matter $_{2.5}$) (69).

2.6 Rationale of the study

Previous studies on methylation levels of *F2RL3* and *AHRR,* and lung cancer risk have focussed on a few CpG sites within each gene. These studies thus assumed that methylation levels in those few preselected sites are representative of regional methylation levels and hence, can serve as proxies for gene expression. Compared to looking at a single CpG site, the examination of multiple CpG sites could potentially provide a more complete measure of methylation. As such, the study of the regional methylation pattern within *F2RL3* and *AHRR* by looking at multiple CpG sites is likely a better representation of their gene expression. A compelling avenue is the investigation of the methylation of multiple CpG sites that are not solely confined to the promoter region within the two genes as not only methylation in promoter region can determine gene expression (70–73).

Chapter 3. Objectives

3.1 Study objective and hypothesis

The objective of this thesis was to measure the associations between regional DNA methylation patterns within *F2RL3* and *AHRR* genes, and lung cancer risk

It was hypothesized that looking at regional methylation patterns is a better representation than the use of a single CpG site of the association between *F2RL3* and *AHRR* methylation, and lung cancer risk. It is also hypothesized that aberrant methylation level (i.e., hypermethylation) of the two genes of interest is inversely associated with lung cancer risk.

Chapter 4. Methods

4.1 Study Design

A case-control study with cumulative sampling was nested in the CARTaGENE study. Started in 2009, CARTaGENE is the largest ongoing prospective cohort study in Quebec, Canada. Briefly, it is composed of men and women, aged between 40 and 69 at baseline, residing in six of the metropolitan areas of Quebec (Montreal, Sherbrooke, Quebec, Saguenay, Gatineau, and Trois-Rivières) (74). Participants were selected through random stratified sampling and were aimed to be broadly representative of the Quebec population density from the 2006 Census and the provincial health insurance registry files [Régie de l'assurance maladie du Québec (RAMQ)]. Participants were enrolled throughout two phases of recruitment: Phase A (from August 2009 to October 2010) and Phase B (from December 2012 and February 2015). Information packages on the study were first mailed to potential participants, followed by an initial contact by telephone through a call center at the RAMQ to enroll them and invite them for an interview and a physical assessment at one of the clinical assessment sites. Whole blood samples, among other biospecimens, were collected at baseline for each participant during the initial interview and stored pending further analysis at Genome Quebec and Saguenay hospital/ECOGENE-21 Biobank. Follow-up of participants was conducted via occasional web-based questionnaires and linkage to governmental health administrative databases, such as the RAMQ (74).

In our case-control study with cumulative sampling, cases (N=187) included all incident lung cancer cases in the CARTaGENE cohort which occurred during the follow-up period of 2009-2015 and who had provided a blood sample. They were identified via linkage with the RAMQ and the Québec cancer registry. Cases were defined according to the codes 1622-1625, 1682, and 1629 from the International Classification of Diseases for Oncology, Tenth Edition (ICD-10). Controls

(N=378) were randomly sampled at the end of the follow-up period (2015), at a ratio of 2:1, among participants who did not develop lung cancer and had isolated DNA samples from blood. They were frequency-matched to cases by sex, age (5 years interval), and phase of blood collection (Phase A: 2009-2011; Phase B: 2012-2015).

4.2 Quantification of methylation

DNA extraction from whole blood samples was carried out by the Biobanque Génome Québec (Chicoutimi) and stored at -80°C. Quantification of DNA methylation of the *F2RL3* and *AHRR* genes was then performed at the CHU Sainte-Justine and Genome Quebec Integrated Centre for Pediatric Clinical Genomic. For each participant, 1 μg of isolated DNA was bisulfite converted using the EZ DNA Methylation-Gold kit (ZymoResearch) and stored at -80°C pending DNA methylation quantification. This bisulfite conversion step deaminates the unmethylated cytosine to uracil, allowing the detection of methylation patterns by selectively differentiating methylated and unmethylated cytosine (3).

Sequenome EpiTYPER® technology was then used to quantify DNA methylation levels of *F2RL3* and *AHRR*. EpiTYPER® is a validated and reproducible high-throughput mass spectrometry-based method to quantify DNA methylation of multiple CpG sites within genomic regions of 100-600 bps(75,76). In brief, bisulfited converted DNA from the regions of interest is amplified by PCR (polymerase chain reaction) and transcribed into a single stranded RNA product, which is further cleaved into specific fragments to be separated through mass spectrometry.

The regions of interest for *F2RL3* spans 4946 base pairs (GRCh37/hg19: chr19:16999071-17004017, negative strand) and 33 599 base pairs (GRCh37/hg19: chr5:367471-401070, positive strand) for *AHRR*. Short single-stranded DNA sequences used in the initiation of DNA synthesis in PCR, also known as PCR primers, were designed for both *F2RL3* and *AHRR*. The primers were

devised to target CpG sites based on the findings of Fasanelli *et al*. (4), proximity to CpG islands or CpG island shores, transcription factor binding sites, DNAse (an enzyme which cleaves DNA) hypersensitive sites (77–79), and H3K27Ac marks suggestive of the presence of an active regulatory domain within each gene (UCSC Genome Browser, http://genome.ucsc.edu/). Additional factors such as their proximity to CpG islands or CpG island shores, transcription factor binding sites, deoxyribonuclease hypersensitive sites, and H3K27Ac marks suggestive of the presence of an active regulatory domain within each gene were considered during their selection (UCSC Genome Browser, http://genome.ucsc.edu/). Seven and 72 CpG sites were selected within the regions of interest of *F2RL3* and *AHRR*, and separated into one and six different assays, respectively.

The RNA products from PCR were specifically cleaved with RNase A into fragments. Mass spectrometry allows differentiation of the fragments containing a methylated CpG dinucleotide, which are 16 Da heavier, resulting in a shift in the corresponding peaks in the mass spectrum. The surface area of each peak is a measure of the number of its relative fragment present in the assay. The methylation ratio is then calculated by dividing the surface area of the peak associated with methylated fragments by the surface area of its corresponding peaks representing all fragments, both methylated and unmethylated. A methylation ratio of 0 or 1 indicated, respectively, a fully unmethylated or methylated CpG site (75). For each assay, 25 ng of bisulfite-converted DNA was used to quantify methylation ratios within CpG units, which is defined as either an individual CpG site or aggregates of multiple CpG sites, located within each assay. In the case of aggregates of multiple CpG sites, the methylation ratio was assigned to each CpG site constituting the unit. Reliability of the data was assessed through estimated coefficient of variation

(CV) between-plates (4.65%) and between-fragments (4.16%) based on two high methylated human DNA quality controls manufactured by EpigenDx that were included on each plate.

4.3 Conceptualization and parametrization of variables

4.3.1 DNA methylation

In the main analysis, the methylation levels of all individual informative CpG sites in the *F2RL3* and *AHRR* genes were conceptualized via two approaches: by calculating an average measure of methylation across all CpG sites of each gene and by looking at each CpG site, individually. All DNA methylation measures were parametrized both as a standardized continuous variable and as a categorical variable, where the quartile distribution among controls was used to determine the categories. As previously mentioned, DNA methylation was measured as a ratio, where 0 represents a fully unmethylated CpG site. Given that DNA methylation is represented at the ratio level, and thus has no natural metric, parametrization as a standardized continuous variable for DNA methylation was considered. Precisely, when looking at an OR calculated from an unstandardized coefficient, we can interpret it as the effect that a one-unit difference in the independent variable has on the dependent variable. However, if there is no unit of measurement, a change of one unit may not hold any meaningful interpretation. Standardization of DNA methylation thus allows for ease of interpretation by transforming the "unit-free" DNA methylation ratio into a variable measured in standard variation units (80). Furthermore, using DNA methylation as a standardized variable, particularly when the ratio of DNA methylation lacks a natural metric, guarantees that the magnitude of change in DNA methylation is sufficient to produce a significant effect on the dependent variable (i.e. the outcome), assuming that there is a relationship between the two. The core of this reasoning lies in Chebyshev's inequality theory,

which states that, for any distribution, at least 75% of all values (i.e. cases) are within two standard deviations (80).

The decision to use a standardized continuous variable for analyzing the relationship between DNA methylation and lung cancer risk is based on the assumption that this relationship is linear. If the association is linear, it may not necessarily be appropriate to look at it categorically since it can lead to a loss of information and statistical power. However, if no linear trend was found across the categories when testing for trend, categorical measures, such as quartiles, were also considered. In such cases, the lowest quartile, representing the lowest methylation category, was chosen as the reference group.

4.3.2 Additional covariates

The controls in this study were frequency-matched to cases by age, sex, and phase of blood sampling. The latter was done to match cases and controls on their timing of entry in the cohort thus removing the potential of varying lengths of storage time of biosamples on affecting DNA methylation measures. Additional potential confounders were identified through a comprehensive literature review on predictors of DNA methylation levels and lung cancer on PubMed, Ovid, and Clarivate Web of Science platforms, which includes EMBASE and MEDLINE databases. Potential confounders were identified using directed acyclic graphs (DAG) (Figure 4.1) to construct minimally sufficient sets to estimate the total effect of DNA methylation on lung cancer. Identified and retained confounders of lung cancer included body mass index (BMI), fruit and vegetable consumption, ethnicity, and smoking (11,27,81,82).

Age and sex: Age and sex were self-reported by the participants through the questionnaire. Age was treated as a continuous variable and sex as a dichotomous discrete variable.

Phase of blood sampling: Phase of blood sampling was treated as a dichotomous discrete variable.

BMI: BMI was calculated as a ratio of weight to the square of height (kg/m2) from the self-reported measures in the baseline questionnaire. BMI was treated as a categorical variable, where the categories were based on the cut points used by the National Institutes of Health (1994). Specifically, a BMI of less than 18.5 kg/m2 was considered "underweight", between 18.5 and 24.9 kg/m2 was considered "normal", between 25.0 and 29.9 was considered "overweight" and a BMI of greater than 30 was considered "obese". The "obese" category grouped the "moderately obese", "severely obese" and "very severely obese" subcategories (N=103, 28, and 13, respectively) given the small cell size of the two latter subcategories. Similarly, the "underweight" category was grouped with the "normal" category due to its small size (N=3). In brief, BMI was represented as three categories, "normal" (N=177), "overweight" (N=200), and obese (N=144).

Fruit and vegetable consumption: Fruits and vegetable consumption was self-reported by the participants through the questionnaire. It was treated as a categorical variable, with three categories representing the number of fruit or vegetable consumed on an average per day: 0-3, 4-6, 7 or more. The variable was initially categorized into seven categories in the questionnaire: none (N=6), 1 (N=13), 2-3 (N=91), 4-6 (N=216), 7-10 (N=169), 11 or more (N=32). However, due to their small size, the none, 1, and 2-3 categories were grouped together and so were the 7-10 and 11 or more categories.

Ethnicity: Participants were asked to self-report their ethnicity in the questionnaire by indicating which category they belonged to: White, Black, Arab, American-Latino/Hispanic, Filipino, South Asian, Occidental Asian, East Asian, Jewish, South East Asian, Aboriginal, and Others. Ethnicity was treated as a categorical variable with two categories, where participants who indicated that they were White were classified as "Caucasian", and all the other participants who specified other ethnicities were classified as "Others" due to the small cell size (N=17).

Smoking: Smoking was conceptualized as a cumulative smoking index (CSI), which was developed by Hoffmann *et al.*(83). It allows to consolidate the multiple metrics on smoking history and behavior. Information on current smoking status of all participants, and for the ones having smoked at least 100 cigarettes in their lifetime (i.e., age at initiation and cessation, and average number of cigarettes smoked per week) was collected through the questionnaire. CSI, a reliable mathematical and continuous measure of smoking history and behaviour, was derived for each participant and parameterized as a standardized continuous variable:

$$\text{CSI} = (1\text{-}0.5^{dur/\tau})\,(0.5^{tsc/\tau})\,\ln(int + 1)$$

Where *dur* is the duration of smoking, *tsc* is the time since cessation, τ is the biological half-life of tobacco carcinogens, and *int* is the average daily amount smoked in cigarettes. Never smokers and participants who smoked less than 100 cigarettes in their lifetime were attributed a CSI of 0.

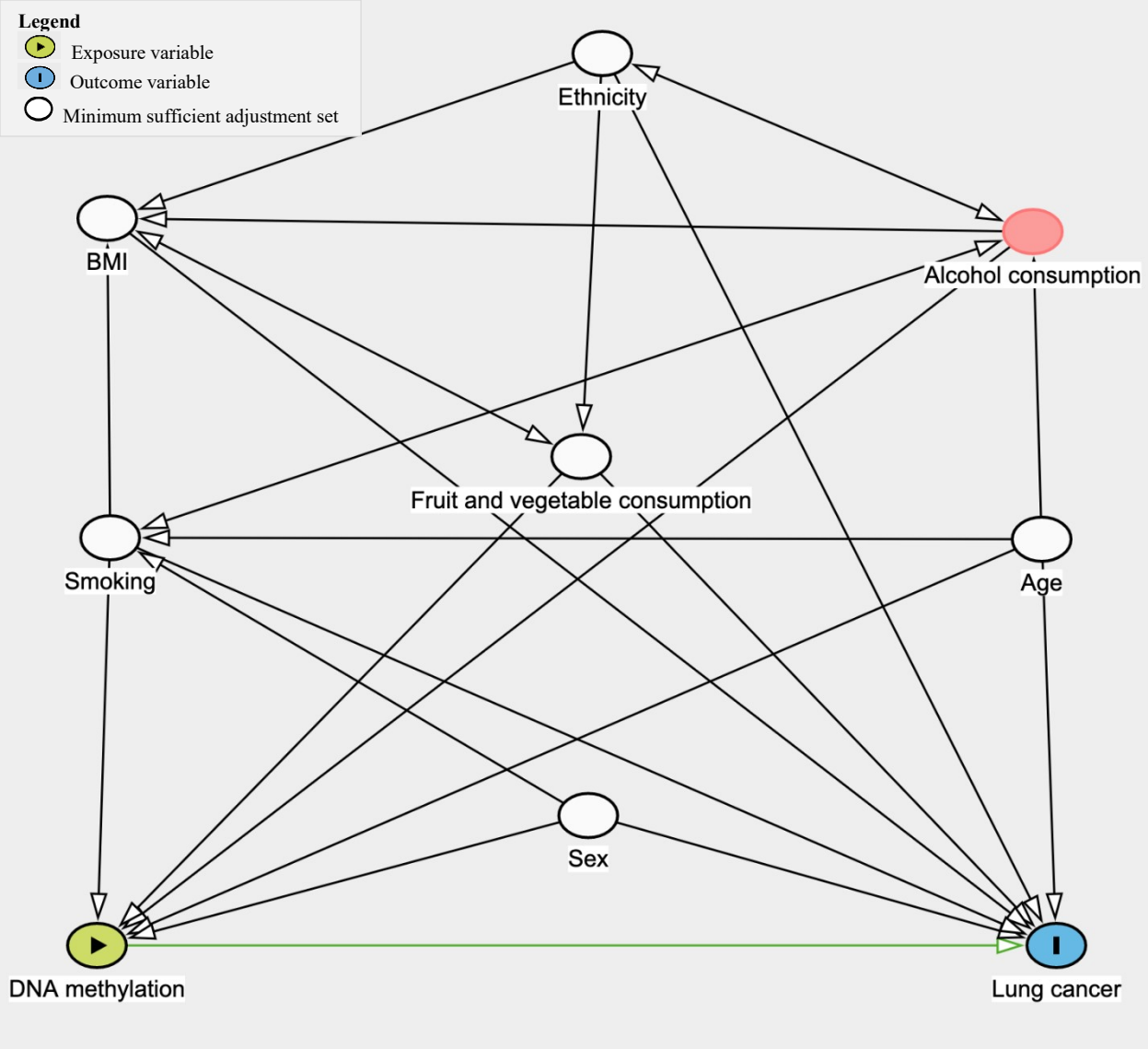

Figure 4.1. Directed acyclic graph of the association between DNA methylation and lung cancer

4.4 Statistical analysis

All statistical analyses were conducted in R, version 4.0.3, (R Core Team, 2020) with the package *dplyr*.

4.4.1 Data cleaning

The cleaning of methylation data was carried out as per Ho *et al*. (76) and summarized in Figure 5.1. Methylation ratios defined as 'unreliable' by Sequenom EpiTYPER® were excluded (N=30 CpG sites for *AHRR* excluded). In brief, fragments overlapping in mass and fragments with a common or uncommon single nucleotide polymorphism(s) (SNP) present were considered unreliable. The former is due to the inability to associate the methylation signal to a unique CpG site as it shares a similar mass to another fragment. As for the latter, the presence of a SNP in the fragment can result in a biased PCR amplification or a bi-modal or tri-modal peak in the spectrum, thus affecting the reliability of those fragments. Fragments with a mass too low or too high for the mass spectrometer to read were also excluded (75). To ensure that only methylation ratios with meaningful differences were included in the calculation of average methylation measures, CpG sites with a s.d. lower than 0.02 were excluded (N=6 CpG sites excluded for *AHRR*). Missing methylation data were assigned the sex-specific mean methylation ratios of their respective CpG site. However, participants with more than 10% of missing methylation data were excluded to ensure that participants in the analytical sample only had a small proportion of missing values imputed (N=8 cases and 26 controls excluded). Similarly, CpG sites with more than 25% of missing participant data were also excluded (N=3 CpG sites excluded for *AHRR*). The remaining CpG sites, which were not defined as 'unreliable' by Sequenom EpiTYPER® criteria and had a s.d. ≥ 0.2 and <10% of missing data, were considered and defined as informative. A total of seven and 33 out of the initial seven and 72 CpG sites for *F2RL3* and *AHRR*, respectively, were retained for 179 cases and 352 controls.

4.4.2 Statistical modeling

Separate multivariate unconditional logistic regression models were used to estimate ORs and 95% CIs for lung cancer risk associated with methylation of the *F2RL3* and *AHRR* genes. The minimally-adjusted model included the variables used for frequency-matching (sex, age, and phase of blood sampling). BMI, fruit and vegetable consumption, ethnicity, and smoking were the potential confounders identified through our DAG and were included in the fully-adjusted model, which also included the variables adjusted for in the minimally-adjusted model (Figure 1). The linearity across the DNA methylation quartiles was assessed via a test for trend. Specifically, the median value of each quartile was treated as a score. This continuous variable was then treated as the independent variable in the logistic models to determine *P*-value for trend across categories, where the null hypothesis indicates no linear trend across the categories (84). Measures with a *P*-value for trend greater than 0.10 in the fully-adjusted model, and which upon visual inspection showed a linear relationship, were considered as linear and represented as standardized continuous measures. Otherwise, categorical representations were also used.

The Benjamini-Hochberg (BH) adjustment was applied to account for the potential inflation of the type I error rate due to the multiple testing in the main analysis (i.e. testing of the association between lung cancer and each individual CpG sites, and for both the minimal- and fully-adjusted models). The BH method was preferred over the Bonferroni correction as it retains statistical power even when a considerable number of tests are done, a situation where the Bonferroni correction is known to be particularly conservative (80, 81).

4.5 Supplementary analysis

Principal component analysis (PCA) was an additional approach considered for the conceptualization of DNA methylation. This was done with the aim to better capture regional

methylation patterns within the two genes and facilitate interpretation, by reducing the number of dimensions in the dataset while minimizing information loss. Briefly, PCA is an approach that reduces the dimensions of the data set, while preserving as much information as possible through the creation of "new" independent variables, also called principal components (PCs) (87). PCs of each gene were obtained, for a maximum of seven and 33 PCs for *F2RL3* and *AHRR*, respectively, using unsupervised PCA, a pattern derivation and data-reduction approach technique (87). The adequate number of PCs to retain was determined based on a combination of multiple criteria: an eigenvalue of ≥1, at least 65-80% of cumulative variance explained, and each PC should account for ≥5% of the variation. Since the data in the unsupervised PCA is standardized, in other words, centered and scaled, the s.d. is also the eigenvalue of the PC. These criteria ensured that the retained PCs explained an adequate proportion of the variation in DNA methylation of each gene, while also allowing to effectively reduce the number of dimensions in the data. In total, two and five PCs were retained out of the initial seven and 33 PCs for *F2RL3* and *AHRR*, respectively. The PCs were parametrized in the same way as the two other approaches (i.e., as a standardized continuous variable and as a categorical variable based on the quartile distribution of controls). Unconditional logistic regression was used to estimate ORs and 95% CI of the retained PCs for lung cancer risk, for both the minimally- and fully-adjusted models.

4.6 Ethical considerations

This study is part of an ongoing project approved on the 27th February 2019 by the Comité d'éthique de la recherche (CER) of Centre hospitalier de l'Université de Montréal (CHUM) co-led by Dr. Vikki Ho and Dr. Anita Koushik. Access to the CARTaGENE database was granted. The Canadian Institute of Health Research (CIHR) supported and funded this project (FRN#162502, 2019).

Chapter 5. Manuscript

This manuscript was written in accordance with the instructions for authors provided by Cancer Epidemiology, Biomarkers & Prevention, a peer-reviewed journal.

Title:
Aberrant DNA methylation of the *F2RL3* and *AHRR* genes and lung cancer risk

Authors:
Alice Nguyen[1,2], Anita Koushik [1,2], Laura Pelland-St-Pierre[1,2], Michael Pham[1,2], Romain Pasquet[1,2], Sherryl Taylor[3], Delphine Bosson-Rieutort[4], Jack Siemiatycki[1,2], Vikki Ho[1,2]

Affiliations

1. Carrefour de l'innovation, Université de Montréal Hospital Research Centre (CRCHUM), Montreal, Québec, Canada
2. Department of Social and Preventative Medicine, Université de Montréal, Montreal, Québec, Canada
3. Department of Medical Genetics, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada
4. Department of Health Management, Evaluation and Policy, University of Montreal, Montreal, Quebec, Canada

Correspondence to:
Dr. Vikki Ho
Université de Montréal Hospital Research Centre (CRCHUM),
850 Saint-Denis Street, 3rd Floor, S03-424
Montreal, Quebec H2X 0A9, Canada
E-mail: vikki.ho@umontreal.ca
Tel: 514-890-8000 ext. 31522
Fax: 514-412-7018

Abstract

Background: The study of biomarkers has the potential to inform on underlying mechanisms in lung cancer etiology. In this study, we investigated DNA methylation in the *F2RL3* and *AHRR* genes, and lung cancer risk.

Methods: A case-control study with cumulative sampling was nested in the CARTaGENE cohort. Cases (N=187) consisted of all participants diagnosed with incident lung cancer from baseline to 2015 and who had provided a blood sample; controls (N=378) were sampled at a ratio of 2:1 with frequency-matching by age, sex, and timing of blood sampling. Sequenom EpiTYPER® was used to quantify methylation levels in seven and 33 CpG sites of *F2RL3* and *AHRR,* respectively. DNA methylation ratios across all individual CpG sites and as an average measure were parametrized both as continuous and categorical variables. Unconditional multivariable logistic regression was used to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for lung cancer associated with *F2RL3* and *AHRR* methylation while controlling for confounders identified using directed acyclic graphs (age, sex, timing of blood sampling, ethnicity, BMI, and fruits and vegetable consumption).

Results: A strong inverse relationship between average DNA methylation levels and lung cancer was observed for both *F2RL3* (OR per standard deviation (s.d.) in methylation change = 0.65, 95% CI: 0.53-0.80) and *AHRR* (OR per s.d. in methylation change = 0.66, 95% CI: 0.53-0.80). Similarly, individual CpG sites showed ORs (per s.d. in methylation change) ranging from 0.61-0.70 for six out of the seven CpG sites of *F2RL3* and from 0.57-0.79 for 17 out of 33 CpG sites of *AHRR.* The remaining CpG sites within *F2RL3* and *AHRR* showed no association with lung cancer risk, except for one CpG site within *AHRR* (chr5:369774) which had an OR of 1.25 (95% CI: 1.02-1.54).

Conclusion: These findings support the role of epigenetic mechanisms in lung cancer etiology.

Keywords: Lung cancer, epigenetics, DNA methylation, aryl-hydrocarbon receptor repressor, coagulation factor II (thrombin) receptor-like 3

Introduction

Prevention is key in the fight against lung cancer as it continues to be the leading cause of cancer mortality despite progress in treatment (1). Although smoking is the principal known risk factor for lung cancer, only a proportion of smokers develop the disease and approximately 10-20% of incident cases occur in never-smokers (2). A better understanding of underlying mechanisms is, therefore, necessary to further progress in prevention initiatives. The study of intermediate endpoints has enormous research potential to further knowledge on the causal mechanism of lung cancer (3) and potentially aid in future screening endeavours. However, their relation to the health outcome must be established in order to be effectively used as a biomarker in population health studies.

The use of DNA methylation as an intermediate event in lung cancer etiology shows promise (4). DNA methylation is an epigenetic process where a methyl group is transferred to sites in the DNA where a cytosine residue is followed by a guanine residue, which are called CpG sites (5). Regions rich in CpG sites are referred to as CpG islands, and those that are 2kb away from the promoter region are referred to as CpG island shores (6). Global genome hypomethylation, and gene-specific hyper- or hypomethylation of CpG islands in promoters of particular genes are two forms of aberrant DNA methylation implicated in lung cancer etiology. The former has been mostly associated with lung cancer progression, a later event in cancer development, while the latter has been linked to early events in lung cancer etiology (4).

An epigenome-wide association study (EWAS) by Fasanelli *et al.* of four pooled prospective cohort studies reported a strong association between lung cancer risk and hypomethylation in the CpG sites cg03636183 of the *F2RL3* gene (OR per s.d. in methylation change = 0.40, 95% CI: 0.31-0.56) and cg05575921 of the *AHRR* gene (OR per s.d. in methylation

change = 0.37, 95% CI: 0.31-0.54) (7). These findings have also been replicated and validated in another EWAS (8). *F2RL3*, or the coagulation factor II receptor like 3 gene, encodes for the protease-activated receptor-4 (PAR-4) which influences blood coagulation and immune response and is involved in neoplastic diseases (9,10). *AHRR* encodes for the aryl hydrocarbon receptor repressor (AHRR), a recognized tumor suppressor gene involved in the inflammatory response, apoptosis, and cell proliferation (11,12). AHRR has also been linked to smoking as many of its known agonists are chemicals found in tobacco smoke, such as polycyclic aromatic carbons (13,14).

The strong associations observed demonstrate the potential of gene-specific DNA methylation to serve as an early marker in lung cancer etiology. However, previous studies have only focused on one to three CpG sites within those two genes. The study of regional patterns of DNA methylation could have greater predictive power and further inform on the association between DNA methylation and lung cancer risk (15). In this study, we aimed to investigate the association between lung cancer and DNA methylation of the *F2RL3* and *AHRR* genes, measured in seven and 33 CpG sites of each gene, respectively.

Methods

1. Study Design and Population

A case-control study with cumulative sampling was nested in the CARTaGENE study. Briefly, CARTaGENE is the largest ongoing prospective cohort in Quebec, Canada (16,17) comprising men and women, aged between 40 and 69, recruited from six metropolitan areas of Quebec (Montreal, Sherbrooke, Quebec, Saguenay, Gatineau, and Trois-Rivières). At the time of enrollment in 2009, participants came to an assessment center for an interview, donated whole

blood samples, and consented to have their information linked to provincial health databases. For this study, cases (N=187) consisted of all incident lung cancer cases occurring during the follow-up period of 2009-2015 identified via linkage with the Régie de l'assurance-maladie du Québec and Québec cancer registry and defined according to the codes 1622-1625, 1682, and 1629 from the International Classification of Diseases for Oncology, Tenth Edition (ICD-10). Controls (N=378) were randomly sampled at the end of the follow-up period (2015), at a ratio of 2:1, among participants who did not develop lung cancer and had isolated DNA from blood. Controls were frequency-matched to cases by sex, age (5 year-interval), and the phase of blood collection (Phase A: 2009-2011; Phase B: 2012-2015).

2. Gene-specific DNA methylation quantification

DNA isolation from baseline blood samples was conducted at the Biobanque Genome Quebec. Quantification of DNA methylation levels in the *AHRR* and *F2RL3* genes was conducted at the CHU Sainte-Justine and Genome Quebec Integrated Centre for Pediatric Clinical Genomics. For each participant, 1 μg of isolated DNA was bisulfite converted using the EZ DNA Methylation-Gold kit (ZymoResearch). Bisulfite conversion allows for the detection of methylation patterns by selectively deaminating unmethylated cytosine to uracil while leaving methylated cytosine unchanged.

DNA methylation levels were quantified using the Sequenom EpiTYPER® technology, a validated and reproducible high-throughput mass spectrometry-based method (18,19). Primers for the two genes were designed from the promoter region based on the findings of Fasanelli *et al.* (14), proximity to CpG islands or CpG island shores, transcription factor binding sites, DNAse (an enzyme which cleaves DNA) hypersensitive sites (20–22), and H3K27Ac marks suggestive of the presence of an active regulatory domain within each gene (UCSC Genome Browser,

http://genome.ucsc.edu/). The region of interest spans 4,946 base pairs for *F2RL3* (GRCh37/hg19: chr19:16999071-17004017, negative strand) and 33,599 base pairs for *AHRR* (GRCh37/hg19: chr5:367471–401070, positive strand). Thirty and 63 CpG sites were analyzed within the regions of interest of *F2RL3* and *AHRR,* and separated into one and six different assays, respectively. For each assay, 25 ng of bisulfite-converted DNA was used to quantify methylation ratios within CpG units (a unit consists of either an individual CpG site or aggregates of multiple CpG sites) located within each assay. Methylation ratios were calculated by dividing the number of methylated cytosine at a specific CpG site of the gene by the total number of copies of that CpG site in the sample. A methylation ratio of 0 or 1 indicated, respectively, a fully unmethylated or methylated CpG site. A methylation ratio of each CpG unit was then assigned to its corresponding individual CpG site or in the case of an aggregate of multiple CpG sites, to each CpG site constituting the unit. Two highly methylated human DNA quality controls manufactured by EpigenDx were included on each plate. A coefficient of variation (CV) of 4.65% and 4.16% was estimated between-plates and between-fragments, respectively, based on the high methylated DNA quality controls.

3. Statistical Analysis

Figure 1 illustrates the processing of the methylation data as described by Ho *et al.* (23). In brief, unreliable methylation ratios for CpG sites were identified according to Sequenom EpiTYPER specifications, and CpG sites with methylation ratios that had a s.d. lower than 0.02 were excluded. The latter restriction was used to ensure that only methylation ratios with meaningful differences were included in the calculation of average methylation measures. Missing methylation data were assigned the sex-specific mean methylation ratios of their respective CpG site. However, participants with more than 10% of missing methylation data were excluded to

ensure that participants in our analytical sample only had a small proportion of missing values imputed. Similarly, CpG sites with more than 25% of missing methylation data were also excluded. A total of seven and 33 CpG sites for *F2RL3* and *AHRR*, respectively, were retained and considered informative for 179 cases and 352 controls.

Multivariate unconditional logistic regression was used to estimate odds ratios (ORs) and 95% confidence intervals (CIs) for lung cancer risk associated with methylation of the *F2RL3* and *AHRR* genes, in separate models. For each gene, we considered both the average methylation level across all informative CpG sites and the methylation level of individual informative CpG sites. All DNA methylation measures were parametrized both as a standardized continuous variable and as a categorical variable. The categories were determined according to the quartile distribution of the methylation measures among controls, where the lowest methylation category was used as the reference group. Linear relationships were identified by a visual evaluation of the dose-response pattern represented categorically as well as the assessment of linear trend across categories. *P*-value for trend across categories was estimated by assigning the mid-point of each category and treating them as a score which was then included in the logistic models (24). CpG sites for which the *P*-trend was less than 0.10, and for which upon visual inspection showed a linear relationship with lung cancer risk were considered as linear and represented as standardized continuous measures. Otherwise, categorical representations were also shown.

Minimally-adjusted models included the frequency-matching factors (sex, age with 5 years intervals, timing of blood sampling). Using a directed acyclic graph (DAG), body mass index (BMI), fruit and vegetable consumption, ethnicity, and smoking were identified as potential confounders and included in the fully-adjusted model. To account for the potential inflation of the

type I error rate as a result of multiple testing in our main analysis, the Benjamini-Hochberg procedure was applied to control for the false discovery rate (25).

Results

In this study, the mean (s.d.) age of participants at recruitment was 58 (7) years, and they were mainly White (95%), and overweight (38%) (Table 1). Relative to controls (13%), cases were more likely to be current smokers (33%).

Table 2 shows the estimated ORs for the association between average DNA methylation levels of *F2RL3* and *AHRR* and lung cancer. Higher average DNA methylation in both genes was associated with a lower lung cancer risk in the fully-adjusted model (*F2RL3*: OR per s.d. of methylation change = 0.65, 95% CI: 0.53-0.80; *AHRR*: OR per s.d. of methylation change = 0.66, 95% CI: 0.53-0.80), indicating that hypomethylation of these two genes was associated with cancer. Similar associations were also observed in the minimally-adjusted models.

The associations found for methylation levels of individual CpG sites in the *F2RL3* and *AHRR* genes and lung cancer risk are summarized in Table 3 and 4, respectively. A total of six out of seven CpG sites for *F2RL3* (range of OR per s.d. of methylation change: 0.61-0.70) and 17 out of 33 CpG sites for *AHRR* (range of OR per s.d. of methylation change: 0.57-0.79) showed an inverse association with lung cancer risk in the fully-adjusted model. Null associations were found for one out of seven CpG sites for *F2RL3* (OR per s.d. of methylation change = 1.08, 95% CI: 0.89-1.31) and 15 out of 33 CpG sites for *AHRR* (range of OR per s.d. of methylation change: 0.84-1.18). One CpG site in the *AHRR* gene (chr5:369774) had an OR of 1.25 (95% CI: 1.02-1.54), denoting hypermethylation in cancer cases relative to controls.

Discussion

Epigenetics is a promising field to further research on the still elusive mechanisms underlying lung carcinogenesis. We believe that our study is the first to measure regional methylation patterns of *F2RL3* and *AHRR* in association with lung cancer, by looking at more CpG sites than previous studies. Across the majority of CpG sites, methylation levels in both genes were inversely associated with lung cancer suggesting hypomethylation of the two genes amongst lung cancer cases relative to controls. These associations are in concordance with an EWAS study of four prospective cohorts which previously described similar inverse associations between lung cancer risk and the methylation levels for the CpG sites, cg03636183 (also referred to as chr19:17000585) within *F2RL3,* and cg05575921 and cg21161138 (also referred to as chr5:373378 and ch5:399360, respectively) within *AHRR* (7). The same research lab validated these associations for cg03636183 in *F2RL3* and cg05575921 and cg23916896 in *AHRR*, the latter being an additional CpG site described in their study, in further analyses adding an additional cohort to the four studies included in their previous EWAS (8).

Our results for *F2RL3* methylation and lung cancer are also in concordance with a prospective study that considered three CpG sites, including cg03636183, within the same region of *F2RL3* and lung cancer (9). However, our findings and that of others contradict another longitudinal study that reported null associations between the methylation level of cg05575921 for *AHRR* and lung cancer risk (26).

Our findings of hypomethylation in *F2RL3* and *AHRR* in lung cancer cases relative to controls are consistent with the current understanding of the mechanisms of both genes. *F2RL3* encodes for PAR-4, a thrombin receptor, which is found to be involved in the process of blood coagulation and its disruption has often been described in lung cancer (27). However, methylation

level of one CpG site out of seven within *F2RL3* (chr19:17000567) was not associated with lung cancer; further research may need to be conducted explain this result as the function and mechanism of PAR-4 have not been fully elucidated.

The observation in this study that hypermethylation of *AHRR* is inversely associated with lung cancer risk is consistent with its role in the AhR pathway (12). Hypomethylation of *AHRR*, and thus the overexpression of AHRR, would disrupt the AhR pathway, which is crucial to the metabolism of environmental toxicants in the body (11,12,28). Its disruption through the overexpression of AHRR could result in the harmful accumulation of such toxicants. One of 33 CpG sites for *AHRR* supported a positive association between hypermethylation and an increased lung cancer risk. This is plausible as *AHRR* has also been shown to be involved in the inflammatory response (12), and its disruption through the hypermethylation of *AHRR*, thus its downregulation, could therefore result in the increased risk for cancer observed.

This study represents the first investigation into the correlation between lung cancer and the methylation of multiple CpG sites in both the *F2RL3* and *AHRR* genes. We hypothesized that their methylation levels could likely be more representative of gene expression than a single CpG site, and thus of the association between *F2RL3* and *AHRR* and lung cancer. Nonetheless, the consistent associations that we observed for the average methylation and the majority of individual CpG sites for *F2RL3* and *AHRR* support that the methylation of the well-studied CpG sites cg03636183 for *F2RL3* and cg05575921 for *AHRR* site can be representative of the methylation level of the other CpG sites located within the same gene.

Previous studies have used microarray chips to quantify methylation within *F2RL3* and *AHRR*. While that technology permits the investigation of epigenomic-wide data, it does not allow for the examination of CpG sites other than the predetermined ones on the microarray. As such,

systematic examination of regional methylation patterns via the quantification of multiple CpG sites within a gene was not possible. In our study, quantification of methylation was carried out at a single nucleotide resolution through the use of Sequenome EpiTYPER®, a reliable method (19) as demonstrated by the obtained CVs of 4.65% and 4.16%, estimated between-plate and between-fragment variation, respectively. We believe that our study is the first to examine methylation levels across considerably more CpG sites, relative to previous studies, within the two genes in association with lung cancer risk; in particular for *AHRR* as 33 CpG sites were retained as compared to seven CpG sites for *F2RL3*.

The Quebec Cancer Registry was used to identify cases up until 2010; subsequently, cases after 2010 were identified through linkage with RAMQ [Régie de l'assurance maladie du Québec], an administrative database of health information on Québec citizens. RAMQ has recently been found to underestimate the number of cases in the Québec population (29). Thus it is possible that the some lung cancer cases in CARTaGENE was not included in our study. However, there is no reason to believe that being diagnosed (or not) in the RAMQ as a case is related to exposure status. While we could also consider an alternate scenario whereby an undiagnosed case was selected as a control, this is unlikely as lung cancer is rare and symptoms are severe and thus, people are likely to seek care and thereby get a diagnosis..

DNA methylation was measured in the blood leukocytes from whole blood samples instead of directly in lung tissues. The heterogeneous nature of whole blood samples and the variability of blood cell type composition between individuals could confound the estimated association as it is widely accepted that DNA methylation measurements can differ between different cell types (30). Due to the lack of an external validation set, the proposed algorithm to adjust for cell type distribution within blood samples could not be applied in our study (31). Five EWASs have noted

however that the variation in DNA methylation due to the heterogonous nature of whole blood is relatively small and insignificant when comparing differential methylation patterns (32–37). The consistency and concordance of our results with previous EWASs (7,8) which have adjusted for cell type composition in whole blood lend confidence that the influence of cell composition in whole blood samples on the association observed was minimal.

Conclusion

Compared to previous studies, we interrogated the methylation level of the largest number of CpG sites within *F2RL3* and *AHRR* in relation to lung cancer risk. This present study supports that *F2RL3* and *AHRR* methylation can be informative biomarkers in lung cancer etiology. Given the previously noted association between smoking and *F2RL3* and *AHRR* methylation, further research using a mediation analysis should be done to address the question of whether *F2RL3* and *AHRR* methylation mediates the smoking-lung cancer association.

Acknowledgment

The investigators would like to thank the participants from the CARTaGENE study who made this project possible.

Authors contribution

VH, AK, and JS conceived and designed the study. With support from VH, LPSP, MP, and RP, AN carried out the statistical analyses. ST was in charge of the quantification of the DNA methylation measures. VH, LPSP, and AN contributed to the interpretation of the results. AN wrote the manuscript under the supervision of VH, LPSP, and AK.

Funding

This study was supported and funded by the Canadian Institutes of Health Research (FRN#162502, 2019).

Table 5.1. Baseline characteristics of study population, n(%)

	All (N=531) n (%) or mean (s.d.)	**Cases (N=179) n (%) or mean (s.d.)**	**Controls (N=352) n (%) or mean (s.d.)**
Age at baseline (years), mean (s.d.)	58±7	59±7	59±7
Sex, n (%)			
Male	260 (49)	86 (48)	174 (49)
Female	271 (51)	93 (52)	178 (51)
Phase of blood sampling, n (%)			
1	483 (91)	162 (91)	321 (91)
2	48 (9)	17 (9)	31 (9)
Smoking status, n (%)			
Never smokers	186 (35)	42 (23)	144 (41)
Former smokers	237 (45)	75 (42)	162 (46)
Current daily smoker	22 (4)	8 (4)	14 (4)
Current occasional smoker	84 (16)	52 (29)	32 (9)
Missing	2 (0)	2 (1)	0 (0)
Ethnicity, n (%)			
White	502 (95)	160 (89)	342 (97)
Other	17 (3)	10 (6)	7 (2)
Missing	12 (2)	9 (5)	3 (1)
Body mass index categories, n (%)			
Normal	177 (33)	65 (36)	112 (31)
Overweight	200 (38)	53 (30)	147 (42)
Obese	144 (27)	56 (21)	88 (25)
Missing	10 (2)	5 (3)	5 (1)
Fruits and vegetable consumption, n (%)			
0-3	110 (21)	47 (26)	63 (18)
4-6	216 (41)	73 (41)	143 (41)
7 +	201 (38)	55 (31)	146 (41)
Missing	4 (1)	4 (2)	0 (0)

Table 5.2. Associations between average DNA methylation of the *F2RL3* and *AHRR* genes and the risk of lung cancer

	Minimally-adjusted model [a,]		**Fully-adjusted model**[b]	
	OR for 1 s.d.[c] (95% CI)	*p*	OR for 1 s.d.[c] (95% CI)	*p*
F2RL3	0.58 (0.48-0.70)	<0.001	0.65 (0.53-0.80)	<0.001
AHRR	0.59 (0.48-0.71)	<0.001	0.66 (0.53-0.80)	<0.001

[a]Adjusted for age, sex, and timing of blood sampling
[b]Adjusted for age, sex, timing of blood sampling, ethnicity, BMI, and fruits and vegetable consumption
[c]OR per 1 standard deviation increase in DNA methylation

Table 5.3. Associations between DNA methylation of individual CpG sites within *F2RL3* and the risk of lung cancer

Position	Methylation categories	Minimally-adjusted model[a]			Fully-adjusted model[b]		
		OR for 1 s.d.[c] (95% CI)	***p***	***p*TREND[d]**	**OR for 1 s.d.[c] (95% CI)**	***p***	***p*TREND[d]**
chr19:17000596	Standardized continuous	0.67 (0.54-0.82)	<.001		0.70 (0.56-0.88)	<.001	
chr19:17000585	Standardized continuous	0.56 (0.46-0.68)	<.001		0.63 (0.51-0.77)	<.001	
chr19:17000567	Standardized continuous	1.10 (0.92-1.33)	0.319		1.13 (0.93-1.39)	0.430	
	Q1	Referent		0.811	Referent		0.792
	Q2	0.65 (0.38-1.12)			0.63 (0.35-1.14)		
	Q3	1.19 (0.73-1.93)			1.24 (0.73-2.11)		
	Q4	0.82 (0.49-1.36)			0.94 (0.54-1.65)		
chr19:17000552	Standardized continuous	0.55 (0.45-0.66)	<.001		0.62 (0.50-0.77)	<.001	
chr19:17000517	Standardized continuous	0.59 (0.48-0.71)	<.001		0.66 (0.54-0.81)	<.001	
	Q1	Referent		<.001	Referent		0.102
	Q2	0.10 (0.05-0.21)			0.14 (0.06-0.29)		
	Q3	0.56 (0.35-0.89)			0.75 (0.45-1.25)		
	Q4	0.42 (0.25-0.68)			0.57 (0.33-0.98)		
chr19:17000476	Standardized continuous	0.55 (0.45-0.66)	<.001		0.62 (0.50-0.77)	<.001	
chr19:17000465	Standardized continuous	0.54 (0.44-0.65)	<.001		0.61 (0.49-0.75)	<.001	

[a]Adjusted for age, sex, and timing of blood sampling

[b]Adjusted for age, sex, timing of blood sampling, ethnicity, BMI, fruits and vegetable consumption, and smoking

[c]OR per 1 standard deviation increase in DNA methylation

[d]p-value for trend across categories was calculated by assigning the median of each category as a score and computed by adding the continuous variable to the logistic models

Table 5.4. Associations between DNA methylation of individual CpG sites within *AHRR* and the risk of lung cancer

Position	Methylation categories	Minimally-adjusted model[a]			Fully-adjusted model[b]		
		OR for 1 s.d.[c] (95% CI)	*p*	p_{TREND}[d]	OR for 1 s.d.[c] (95% CI)	*p*	p_{TREND}[d]
chr5:373249	Standardized continuous	0.54 (0.43-0.66)	<.001		0.59 (0.47-0.73)	<.001	
chr5:373251	Standardized continuous	0.54 (0.43-0.66)	<.001		0.59 (0.47-0.73)	<.001	
chr5:373300	Standardized continuous	0.52 (0.42-0.63)	<.001		0.57 (0.46-0.71)	<.001	
chr5:373316	Standardized continuous	0.66 (0.55-0.79)	<.001		0.73 (0.59-0.89)	0.002	
	Q1	Referent		0.006	Referent		0.167
	Q2	0.57 (0.34-0.93)			0.58 (0.34-1.00)		
	Q3	0.63 (0.38-1.02)			0.69 (0.40-1.18)		
	Q4	0.47 (0.28-0.79)			0.68 (0.38-1.20)		
chr5:373378	Standardized continuous	0.58 (0.48-0.70)	<.001		0.64 (0.52-0.79)	<.001	
chr5:373424	Standardized continuous	0.55 (0.45-0.67)	<.001		0.61 (0.49-0.75)	<.001	
chr5:373477	Standardized continuous	0.61 (0.51-0.73)	<.001		0.69 (0.56-0.84)	<.001	
chr5:373491	Standardized continuous	0.59 (0.49-0.71)	<.001		0.66 (0.53-0.81)	<.001	
chr5:373495	Standardized continuous	0.59 (0.49-0.71)	<.001		0.66 (0.53-0.80)	<.001	
chr5:373530	Standardized continuous	0.61 (0.51-0.73)	<.001		0.71 (0.57-0.88)	0.002	
chr5:373610	Standardized continuous	0.89 (0.74-1.07)	0.213		0.94 (0.77-1.15)	0.54	
	Q1	Referent		0.248	Referent		0.691
	Q2	0.59 (0.35-0.99)			0.58 (0.32-1.04)		
	Q3	0.92 (0.57-1.50)			0.96 (0.56-1.66)		
	Q4	0.68 (0.40-1.13)			0.80 (0.46-1.41)		
chr5:368449	Standardized continuous	0.76 (0.63-0.91)	0.003		0.84 (0.69-1.02)	0.088	
chr5:368447	Standardized continuous	0.76 (0.63-0.91)	0.003		0.84 (0.69-1.02)	0.088	
chr5:368430	Standardized continuous	0.88 (0.74-1.05)	0.149		0.94 (0.77-1.14)	0.525	
	Q1	Referent		0.378	Referent		0.947
	Q2	1.58 (0.97-2.58)			1.66 (0.98-2.83)		
	Q3	0.90 (0.53-1.53)			1.03 (0.57-1.85)		
	Q4	0.79 (0.45-1.35)			0.95 (0.52-1.73)		
chr5:368278	Standardized continuous	1.07 (0.89-1.30)	0.468		1.11 (0.91-1.36)	0.319	
chr5:368762	Standardized continuous	1.01 (0.84-1.21)	0.897		0.98 (0.81-1.20)	0.845	

chr5:368805	Standardized continuous	0.76 (0.62-0.91)	0.004		0.77 (0.63-0.94)	0.013	
chr5:368898	Standardized continuous	0.94 (0.78-1.13)	0.516		0.88 (0.71-1.07)	0.197	
	Q1	Referent		0.587	Referent		0.279
	Q2	0.73 (0.42-1.25)			0.69 (0.38-1.24)		
	Q3	1.62 (1.01-2.62)			1.40 (0.84-2.34)		
	Q4	0.62 (0.35-1.08)			0.53 (0.29-0.97)		
chr5:368900	Standardized continuous	0.94 (0.78-1.13)	0.516		0.88 (0.71-1.07)	0.197	
	Q1	Referent		0.587	Referent		0.279
	Q2	0.73 (0.42-1.25)			0.69 (0.38-1.24)		
	Q3	1.62 (1.01-2.62)			1.40 (0.84-2.34)		
	Q4	0.62 (0.35-1.08)			0.53 (0.29-0.97)		
chr5:369774	Standardized continuous	1.21 (1.01-1.46)	0.041		1.25 (1.02-1.54)	0.033	
chr5:369970	Standardized continuous	0.90 (0.75-1.08)	0.271		0.90 (0.74-1.11)	0.321	
	Q1	Referent		0.717	Referent		0.600
	Q2	0.78 (0.45-1.34)			0.70 (0.39-1.27)		
	Q3	1.42 (0.87-2.34)			1.31 (0.76-2.26)		
	Q4	0.92 (0.54-1.55)			0.97 (0.55-1.70)		
chr5:370021	Standardized continuous	1.14 (0.95-1.37)	0.145		1.18 (0.97-1.44)	0.099	
chr5:377325	Standardized continuous	0.83 (0.69-1.00)	0.050		0.78 (0.63-0.96)	0.019	
	Q1	Referent		0.340	Referent		0.140
	Q2	0.57 (0.34-0.95)			0.53 (0.30-0.91)		
	Q3	0.51 (0.30-0.85)			0.48 (0.27-0.83)		
	Q4	0.78 (0.48-1.26)			0.67 (0.39-1.14)		
chr5:377359	Standardized continuous	0.66 (0.53-0.80)	<.001		0.67 (0.53-0.83)	<.001	
chr5:377361	Standardized continuous	0.66 (0.53-0.80)	<.001		0.67 (0.53-0.83)	<.001	
chr5:377438	Standardized continuous	0.78 (0.64-0.95)	0.013		0.78 (0.63-0.96)	0.022	
	Q1	Referent		0.514	Referent		0.477
	Q2	1.70 (1.05-2.79)			1.80 (1.05-3.15)		
	Q3	0.54 (0.29-0.97)			0.53 (0.27-1.02)		
	Q4	1.16 (0.67-2.00)			1.16 (0.63-2.13)		

chr5:377453	Standardized continuous	1.00 (0.83-1.20)	0.989		0.99 (0.82-1.2)	0.927	
	Q1	Referent		0.577	Referent		0.570
	Q2	1.47 (0.90-2.43)			1.28 (0.74-2.21)		
	Q3	0.49 (0.26-0.89)			0.47 (0.24-0.90)		
	Q4	1.41 (0.86-2.34)			1.39 (0.80-2.41)		
chr5:392693	Standardized continuous	1.08 (0.89-1.29)	0.441		1.09 (0.89-1.33)	0.394	
	Q1	Referent		0.591	Referent		0.732
	Q2	1.45 (0.87-2.43)			1.46 (0.84-2.55)		
	Q3	1.31 (0.77-2.24)			1.26 (0.70-2.28)		
	Q4	0.99 (0.57-1.73)			1.03 (0.56-1.89)		
chr5:392704	Standardized continuous	1.00 (0.83-1.20)	0.964		1.06 (0.87-1.29)	0.533	
	Q1	Referent		0.879	Referent		0.691
	Q2	1.00 (0.60-1.66)			0.98 (0.56-1.72)		
	Q3	1.00 (0.60-1.67)			1.06 (0.61-1.86)		
	Q4	0.97 (0.58-1.62)			1.12 (0.63-1.97)		
chr5:392940	Standardized continuous	0.76 (0.63-0.91)	0.004		0.79 (0.64-0.96)	0.021	
	Q1	Referent		0.102	Referent		0.209
	Q2	0.78 (0.48-1.27)			0.80 (0.47-1.36)		
	Q3	0.54 (0.32-0.92)			0.54 (0.30-0.96)		
	Q4	0.76 (0.46-1.24)			0.82 (0.47-1.42)		
chr5:392946	Standardized continuous	0.76 (0.63-0.91)	0.004		0.79 (0.64-0.96)	0.021	
	Q1	Referent		0.102	Referent		0.209
	Q2	0.78 (0.48-2.90)			0.80 (0.47-1.36)		
	Q3	0.54 (0.32-0.92)			0.54 (0.30-0.96)		
	Q4	0.76 (0.46-1.24)			0.82 (0.47-1.42)		
chr5:393073	Standardized continuous	0.78 (0.65-0.94)	0.008		0.84 (0.69-1.02)	0.087	
	Q1	Referent		0.097	Referent		0.527
	Q2	0.76 (0.46-1.23)			0.90 (0.53-1.54)		
	Q3	0.48 (0.28-0.81)			0.58 (0.32-1.05)		
	Q4	0.74 (0.45-1.21)			0.93 (0.54-1.60)		

chr5:393076	Standardized continuous	0.78 (0.65-0.94)	0.008		0.84 (0.69-1.02)	0.087	
	Q1	Referent		0.097	Referent		0.527
	Q2	0.76 (0.46-1.23)			0.90 (0.53-1.54)		
	Q3	0.48 (0.28-0.81)			0.58 (0.32-1.05)		
	Q4	0.74 (0.45-1.21)			0.93 (0.54-1.60)		

[a]Adjusted for age, sex, and timing of blood sampling
[b]Adjusted for age, sex, timing of blood sampling, ethnicity, BMI, fruits and vegetable consumption, and smoking
[c]OR per 1 standard deviation increase in DNA methylation
[d]p-value for trend across categories was calculated by assigning the median of each category as a score and computed by adding the continuous variable to the logistic models

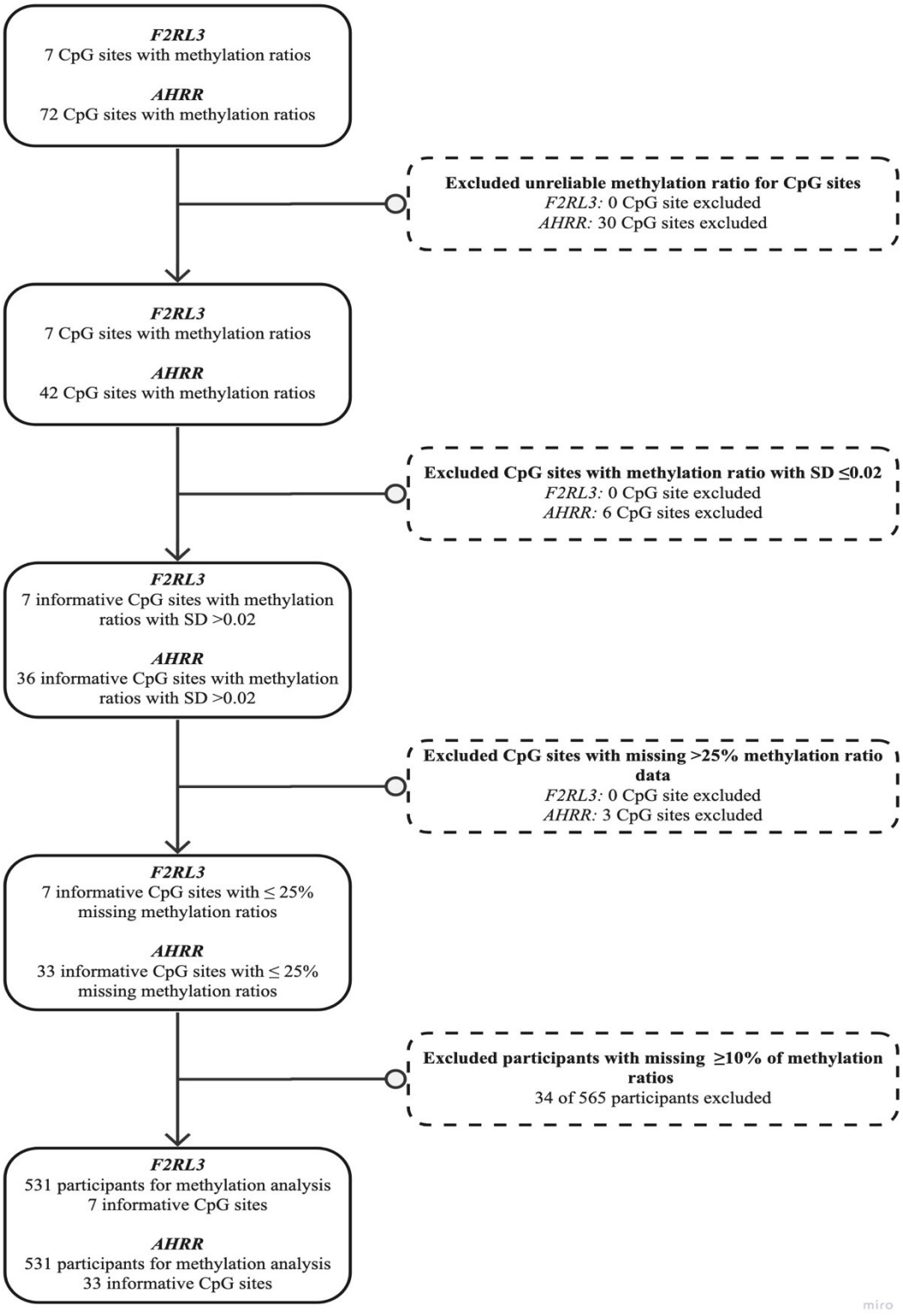

Figure 5.1. Data cleaning of methylation ratios of CpG sites within *F2RL3* and *AHRR* genes.

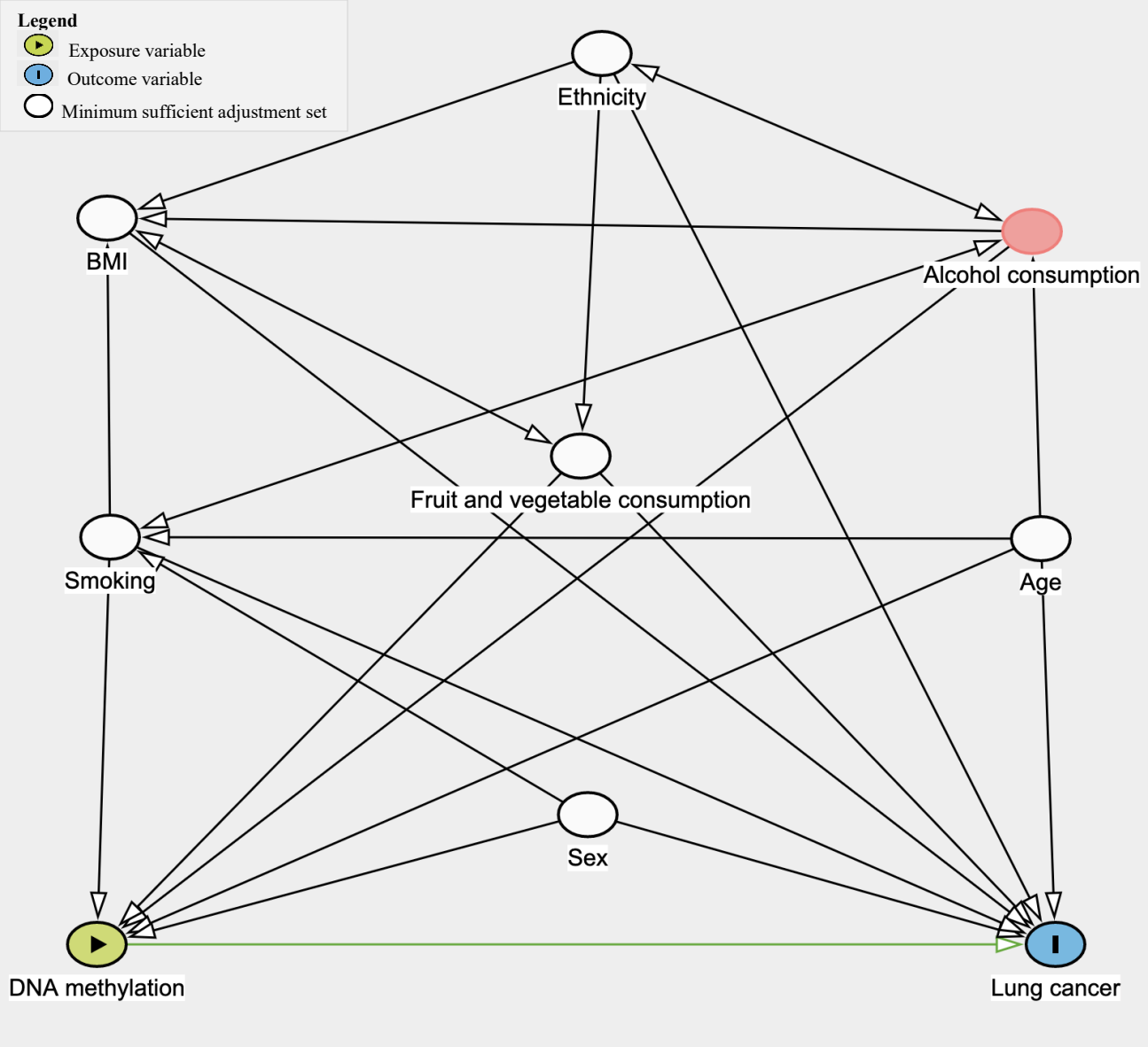

Figure 5.2. Directed acyclic graph of the association between DNA methylation and lung cancer

References

1. World Health Organization. Cancer today [Internet]. Int. Agency Res. Cancer. 2020 [cited 2021 Jul 13]. Available from: http://gco.iarc.fr/today/home

2. Wakelee HA, Chang ET, Gomez SL, Keegan TH, Feskanich D, Clarke CA, et al. Lung cancer incidence in never smokers. J Clin Oncol Off J Am Soc Clin Oncol. 2007;25:472–8.

3. Anglim PP, Alonzo TA, Laird-Offringa IA. DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update. Mol Cancer. 2008;7:81.

4. Brzeziańska E, Dutkowska A, Antczak A. The significance of epigenetic alterations in lung carcinogenesis. Mol Biol Rep. 2013;40:309–25.

5. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. Nature Publishing Group; 2012;13:484–92.

6. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. Nature Publishing Group; 2010;28:1057–68.

7. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6.

8. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung C, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2017;140:50–61.

9. Zhang Y, Yang R, Burwinkel B, Breitling LP, Holleczek B, Schöttker B, et al. F2RL3 methylation in blood DNA is a strong predictor of mortality. Int J Epidemiol. 2014;43:1215–25.

10. Gomides LF, Duarte ID, Ferreira RG, Perez AC, Francischi JN, Klein A. Proteinase-activated receptor-4 plays a major role in the recruitment of neutrophils induced by trypsin or carrageenan during pleurisy in mice. Pharmacology. 2012;89:275–82.

11. Vogel CFA, Haarmann-Stemmann T. The aryl hydrocarbon receptor repressor – More than a simple feedback inhibitor of AhR signaling: Clues for its role in inflammation and cancer. Curr Opin Toxicol. 2017;2:109–19.

12. Zudaire E, Cuesta N, Murty V, Woodson K, Adams L, Gonzalez N, et al. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. J Clin Invest. 2008;118:640–50.

13. Cole JW, Xu H. Aryl Hydrocarbon Receptor Repressor Methylation: A Link Between Smoking and Atherosclerosis. Circ Cardiovasc Genet. 2015;8:640–2.

14. Dertinger SD, Silverstone AE, Gasiewicz TA. Influence of aromatic hydrocarbon receptor-mediated events on the genotoxicity of cigarette smoke condensate. Carcinogenesis. 1998;19:2037–42.

15. Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Nucleic Acids Res. 2017;45:5100–11.

16. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. Oxford Academic; 2013;42:1285–99.

17. Borugian MJ, Robson P, Fortier I, Parker L, McLaughlin J, Knoppers BM, et al. The Canadian Partnership for Tomorrow Project: building a pan-Canadian research platform for disease prevention. CMAJ Can Med Assoc J. 2010;182:1197–201.

18. Ehrich M, Nelson MR, Stanssens P, Zabeau M, Liloglou T, Xinarianos G, et al. Quantitative high-throughput analysis of DNA methylation patterns by base-specific cleavage and mass spectrometry. Proc Natl Acad Sci. National Academy of Sciences; 2005;102:15785–90.

19. Suchiman HED, Slieker RC, Kremer D, Slagboom PE, Heijmans BT, Tobi EW. Design, measurement and processing of region-specific DNA methylation assays: the mass spectrometry-based method EpiTYPER. Front Genet. 2015;6.

20. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. American Association for the Advancement of Science; 2012;337:1190–5.

21. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. Nature Publishing Group; 2011;480:490–5.

22. Kyung Lee M, Armstrong DA, Hazlett HF, Dessaint JA, Mellinger DL, Aridgides DanielS, et al. Exposure to extracellular vesicles from Pseudomonas aeruginosa result in loss of DNA methylation at enhancer and DNase hypersensitive site regions in lung macrophages. Epigenetics. 16:1187–200.

23. Ho V, Ashbury JE, Taylor S, Vanner S, King WD. Quantification of gene-specific methylation of DNMT3B and MTHFR using sequenom EpiTYPER®. Data Brief. 2016;6:39–46.

24. Greenland, Sander. Analysis of polytomous exposures and outcomes, in: K.J.Rothman, S. Greenland (Eds.). Mod Epidemiol. 2nd ed. Philadelphia: Lippincott Williams &Wilkins; 1998. page 301.

25. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. [Royal Statistical Society, Wiley]; 1995;57:289–300.

26. Battram T, Richmond RC, Baglietto L, Haycock PC, Perduca V, Bojesen SE, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol. 2019;48:1493–504.

27. Ferrigno D, Buccheri G, Ricca I. Prognostic significance of blood coagulation tests in lung cancer. Eur Respir J. 2001;17:667–73.

28. Murray IA, Patterson AD, Perdew GH. Aryl hydrocarbon receptor ligands in cancer: friend and foe. Nat Rev Cancer. Nature Publishing Group; 2014;14:801–14.

29. Diop M, Strumpf EC, Datta GD. Measuring colorectal cancer incidence: the performance of an algorithm using administrative health data. BMC Med Res Methodol. 2018;18:38.

30. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA Methylation in Whole Blood: Uses and Challenges. Curr Environ Health Rep. 2015;2:145–54.

31. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

32. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics. 2015;7.

33. Tsaprouni LG, Yang T-P, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9:1382–96.

34. Guida F, Sandanger TM, Castagné 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.

35. Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet. 2013;132:1027–37.

36. Besingi W, Johansson Å. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23:2290–7.

37. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.

Chapter 6. Supplementary results

The objective of this thesis was to examine the association between DNA methylation in the *F2RL3* and *AHRR* genes and lung cancer risk. The main analysis considered average methylation as well as individual methylation measures of CpG sites for each gene in relation to lung cancer risk. An additional conceptualization of DNA methylation levels of *F2RL3* and *AHRR* was carried out using unsupervised PCA, a pattern derivation and data-reduction approach. This was done to facilitate interpretation by reducing the number of dimensions in the dataset while minimizing information loss. This chapter will present the results of the PCA for *F2RL3* and *AHRR*, and its association with lung cancer risk.

6.1 Computation of the principal components of *F2LR3* and *AHRR*

Table 6.1 and 6.2 summarise the variation explained by the seven PCs of *F2RL3* and the first 10 PCs of *AHRR*, respectively. Scree plots were used to identify the number of PCs with eigenvalues >1 for *F2RL3* and *AHRR* (Figure 6.1 and 6.2, respectively). For *F2RL3*, the first two PCs out of seven have eigenvalues of approximately ≥1 and explained 85% of the variation. Eight out of 33 PCs for *AHRR* respect the criteria of eigenvalues of ≥1, accounting for 79% of the variation. Nevertheless, we elected to retain only the first five PCs of *AHRR* (accounting for 67% of the variation) since each PC afterward accounted for less than 5% of the variation.

6.2 Assessing the association of the principal components for *F2RL3* and *AHRR* with lung cancer risk

Table 6.3 and 6.4 present the ORs and 95% CIs for the PCs of *F2RL3* and *AHRR* estimated for both the minimally- and fully-adjusted models. Both PCs for *F2RL3* showed an inverse relationship with lung cancer risk in both the minimally- and fully-adjusted models. For *AHRR*, only the first PC was inversely associated with lung cancer risk (OR per 1 s.d. increase = 0.87,

95%CI: 0.82-0.93). No significant associations were observed between the other four PCs of *AHRR* and lung cancer risk.

Table 6.1 Summary of variation explained by the seven principal components of *F2RL3*

Principal component of *F2RL3*	Standard deviation[a]	Proportion of variance	Cumulative proportion
1	2.25	0.72	0.72
2	0.96	0.13	0.85
3	0.88	0.11	0.96
4	0.40	0.02	0.98
5	0.27	0.01	0.99
6	0.19	0.01	1.00
7	0.00	0.00	1.00

[a] Also referred to as eigenvalues, as data has been centered and scaled (standardized)

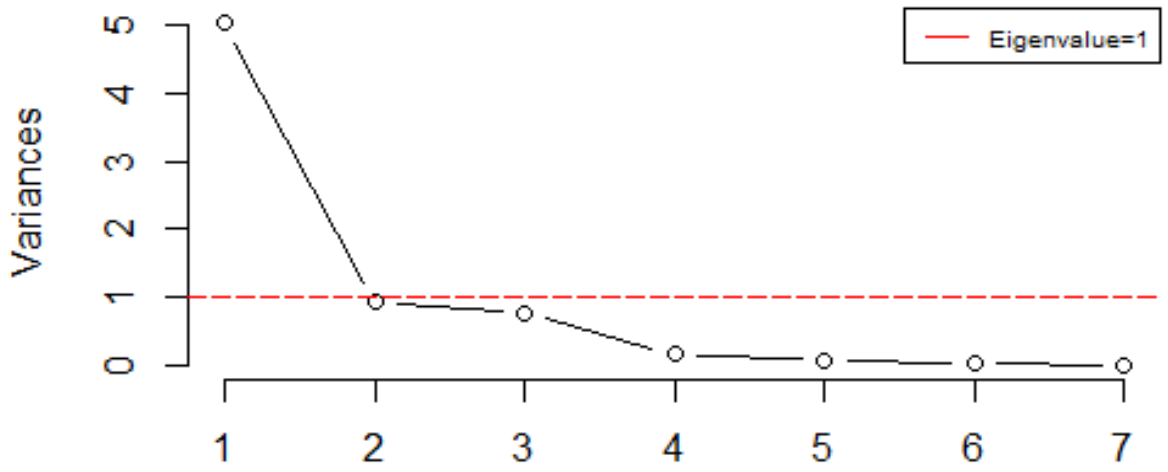

Figure 6.1. Scree plot for the seven principal components of *F2RL3*

Table 6.2 Summary of variation explained by the first 10 principal components of *AHRR*

Principal component of *AHRR*	Standard deviation[a]	Proportion of variance	Cumulation proportion
1	3.40	0.35	0.35
2	1.88	0.11	0.46
3	1.65	0.08	0.54
4	1.48	0.07	0.61
5	1.41	0.06	0.67
6	1.27	0.05	0.72
7	1.16	0.04	0.76
8	1.01	0.03	0.79
9	0.95	0.03	0.81
10	0.94	0.03	0.84

[a] Also referred to as eigenvalues, as data has been centered and scaled (standardized)

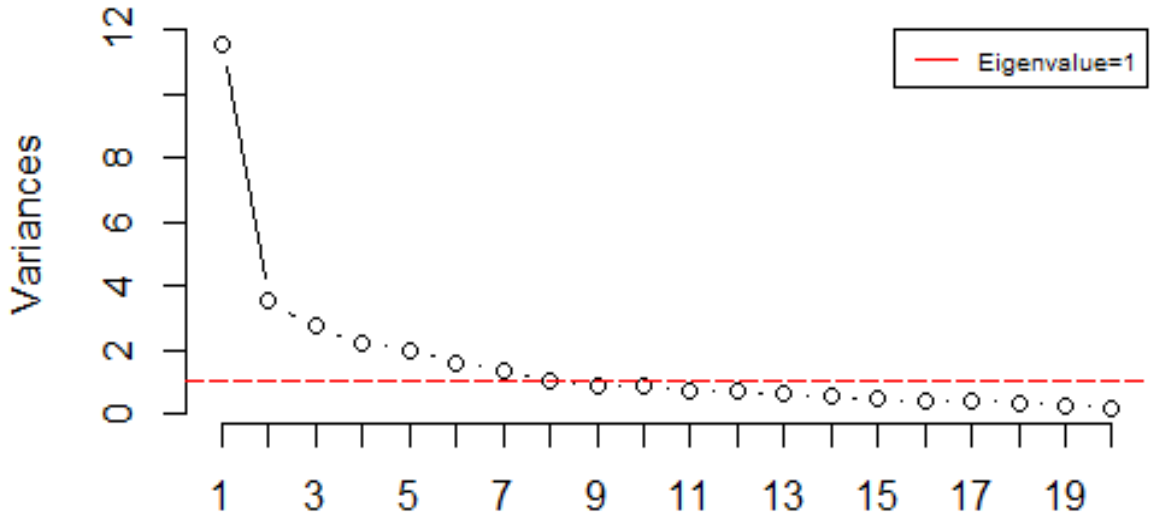

Figure 6.2. Scree plot for the first 20 principal components of *AHRR*

Table 6.3 Associations between DNA methylation of retained principal components of *F2RL3* and the risk of lung cancer

Principal Component	Methylation categories	Minimally-adjusted model[a]		Fully-adjusted model[b]	
		OR for 1 s.d.[c] (95% CI)	***p***	**OR for 1 s.d.[c] (95% CI)**	***p***
1	Continuous	0.77 (0.71-0.84)	<0.001	0.82 (0.74-0.90)	<0.001
2	Continuous	0.70 (0.57-0.85)	<0.001	0.71 (0.56-0.88)	0.003

[a]Adjusted for age, sex, and timing of blood sampling
[b]Adjusted for age, sex, timing of blood sampling, ethnicity, BMI, and fruits and vegetable consumption
[c]OR per 1 standard deviation increase in DNA methylation

Table 6.4 Associations between DNA methylation of retained principal components of *AHRR* and the risk of lung cancer

Principal Component	Methylation categories	Minimally-adjusted model[a]			Fully-adjusted model[b]		
		OR for 1 s.d.[c] (95% CI)	*p*	p_{TREND}[d]	OR for 1 s.d.[c] (95% CI)	*p*	p_{TREND}[d]
1	Continuous	0.85 (0.80-0.89)	<0.001		0.87 (0.82-0.93)	<0.001	
2	Continuous	1.09 (0.99-1.21)	0.082		1.11 (0.99-1.24)	0.073	
3	Continuous	0.93 (0.84-1.05)	0.247		0.94 (0.83-1.05)	0.282	
	Q1	Referent		0.323	Referent		0.436
	Q2	0.51 (0.30-0.86)			0.58 (0.33-1.01)		
	Q3	0.47 (0.28-0.79)			0.55 (0.31-0.97)		
	Q4	0.83 (0.52-1.34)			0.85 (0.50-1.43)		
4	Continuous	0.91 (0.80-1.03)	0.156		0.91 (0.79-1.04)	0.173	
	Q1	Referent		0.361	Referent		0.515
	Q2	1.00 (0.61-1.65)			1.06 (0.61-1.83)		
	Q3	0.95 (0.57-1.58)			1.03 (0.59-1.80)		
	Q4	0.80 (0.47-1.35)			0.84 (0.47-1.50)		
5	Continuous	1.05 (0.92-1.19)	0.483		1.12 (0.98-1.30)	0.109	
	Q1	Referent		0.936	Referent		0.291
	Q2	0.66 (0.38-1.12)			0.76 (0.42-1.37)		
	Q3	0.98 (0.59-1.61)			1.11 (0.64-1.92)		
	Q4	0.95 (0.57-1.56)			1.27 (0.74-2.21)		

[a]Adjusted for age, sex, and timing of blood sampling
[b]Adjusted for age, sex, timing of blood sampling, ethnicity, BMI, fruits and vegetable consumption, and smoking
[c]OR per 1 standard deviation increase in DNA methylation
[d]p-Value for trend across categories was calculated by assigning the median of each category as a score and computed by adding the continuous variable to the logistic models

Chapter 7. Discussion

This project investigated the association between the DNA methylation levels of *F2RL3* and *AHRR,* and lung cancer risk. It made use of the readily available data from CARTaGENE by nesting a case-control study with cumulative sampling in the cohort. The following chapter presents a summary of our results and discusses their contribution to the relevant literature while considering the strengths and limitations of the study.

7.1 Summary of key findings

The results of our main analysis support our hypothesis that there is an association between the methylation levels of *F2RL3* and *AHRR,* and lung cancer risk. Specifically, an inverse relationship between the average methylation levels of all informative CpG sites of *F2RL3* and *AHRR,* and lung cancer risk was observed, indicating hypomethylation of those two genes in lung cancer cases relative to controls. Considering the methylation level of the individual CpG sites, similar associations were observed for six out of the seven CpG sites for *F2RL3* and 17 out of the 33 CpG sites for *AHRR*. Conversely, a positive association with one individual CpG site for *AHRR* and lung cancer risk was observed.

Different conceptualizations of DNA methylation were used in this study. While the use of an average measure of the methylation allows for ease of interpretation, aggregating multiple CpG sites within one measure can result in a loss of information from the individual-level data. On the other hand, looking at all the CpG sites individually, while precise, can complicate the interpretation of our results. A principal component analysis was thus carried out with the aim to reduce the number of dimensions in the dataset while facilitating interpretation and information loss. Two and five PCs were retained for *F2RL3* and *AHRR*, respectively. Both PCs for *F2RL3* showed an inverse association with lung cancer risk, supporting hypomethylation of *F2RL3* in

lung cancer development. An inverse association was noted for only one of the PCs for *AHRR*. The other 4 PCs for *AHRR* showed no association with lung cancer risk. A closer look at the location of each CpG site included in each PC within both *F2RL3* and *AHRR* showed that the CpG sites were all largely dispersed around the gene body. It was therefore difficult to interpret the association of regional methylation of *F2RL3* and *AHRR* with lung cancer risk using PCs as an aggregate measure of regional methylation.

7.2 Comparison with relevant literature

The results from this study support the association between *F2RL3* and *AHRR* methylation levels and lung cancer risk that have been previously reported in other studies (4,62,88,89). Considering individual CpG sites, these prior studies have, however, only focused on very few individual CpG sites within the two genes of interest. An EWAS pooling data from four prospective cohort studies has reported an inverse association with lung cancer risk for only one individual CpG site within *F2RL3* (chr19:17000585, also known as cg03636183) and two within *AHRR* (chr5:373378, also referred to as cg05575921; and ch5:399360, also referred to as cg21161138) (4). The same research lab validated these associations for cg03636183 in *F2RL3,* and cg05575921 and cg23916896 in *AHRR*, the latter being an additional CpG site described in their study, in further analyses adding an additional cohort to the four studies included in their previous EWAS (62). Similarly, a meta-analysis of the same four cohorts replicated similar inverse associations of the same magnitude for cg03636183 in *F2RL3* and cg05575921 in *AHRR* (88). Interestingly, that same study also carried out a two-sample Mendelian randomization on a different prospective cohort, a method which uses genetic variants as instrument variables to investigate whether DNA methylation is on the causal pathway (90,91), and found no evidence that differential methylation of cg03636183 for *F2RL3* and cg05575921 for *AHRR* has potential

causal effects on lung cancer risk (88). The association with lung cancer risk found for the individual CpG sites for *F2RL3* are in agreement with a longitudinal epigenetic study, which looked at three CpG sites within the same region of *F2RL3*, including cg03636183 (89). While the associations described in this study for the individual CpG sites within *AHRR* are in concordance with the previous EWAS studies, one study, which looked at the methylation of cg05575921 for *AHRR* in heavy smokers, found no association of its methylation with lung cancer risk (92). Despite being largely similar in directionality to the associations reported in previous studies, the associations for *F2RL3* and *AHRR* methylation found in our study tend slightly closer to the null than the associations reported in previous studies. For example, previous reported inverse associations ranged between 0.40-0.64 for cg03636183 of *F2RL3* and 0.37-.0.53 for cg05575921 of *AHRR* (4,62,88). However, the ORs found in our study were 0.63 (95% CI: 0.51-0.77) for cg03636183 of *F2RL3* and 0.64 (95% CI: 0.52-0.79) for cg05575921 of *AHRR*.

The inverse associations found for average methylation of *F2RL3* and *AHRR* are also in agreement in magnitude and directionality with the inverse associations found for the individual CpG sites in our studies. As such, they are also consistent with the associations reported for the individual CpG sites reported in the literature. None of the previous studies have looked at average methylation of *F2RL3* and *AHRR* in relation to lung cancer risk. This is probably due to the fact that previous studies were limited by the few numbers of CpG sites examined to consider a different methylation representation.

Indeed, relatively few CpG sites within *F2RL3* and *AHRR* were examined in relation to lung cancer risk in previous studies. Actually, most studies have examined one CpG site within *F2RL3* (cg03636183). Except for the one study which has looked at three CpG sites, including cg03636183, which were however all within the same region of *F2RL3* (89). Similarly, most

studies have only looked at one CpG site within *AHRR* (cg05575921). Only two studies out of the previous studies have looked at one more CpG site in addition to cg05575921, which were 4,575 (62) and 25,982 bps (4) away from cg05575921. While our study covers a slightly smaller range in bps for *AHRR* (from 130 to 19,699 bps away from cg05575921), our study has examined considerably more CpG sites within *F2RL3* and *AHRR* than previous studies, with 7 and 33 CpG sites, respectively.

Our study aimed to further knowledge on the association of the regional methylation patterns within a *F2RL3* and *AHRR* with lung risk by looking at the methylation level of multiple CpG sites. Examining multiple CpG sites within both genes and their varying methylation levels is likely a better representation of gene expression in relation to lung cancer risk than looking at a single CpG site. It was, however, shown in this study that the methylation of the well-studied CpG sites cg03636183 for *F2RL3* and cg05575921 for *AHRR* site can be representative of the methylation level of the other CpG sites located within the same gene given the consistent inverse associations with lung cancer risk observed for the majority of CpG sites in *F2RL3* and *AHRR*.

Additional studies are still needed to support future development to ascertain whether methylation markers can be a tool for lung cancer screening. A recent study using participants from the population-based Copenhagen City Heart Study in Denmark has looked at the inclusion of the *AHRR* methylation to improve eligibility criteria to identify individuals at risk from seven lung cancer screening trials. They have found that the addition of the methylation of cg05575921 in *AHRR* as an eligibility criterion results in a higher specificity for all current screening criteria (93). The potential of the methylation of *F2RL3* and *AHRR* to lower the screening burden warrants further investigation into the use of methylation markers as a biomarker.

7.3 Potential mechanisms of the methylation of *F2RL3* and *AHRR* in lung carcinogenesis

With regards to *F2RL3*, most of our results concord with previous studies, indicating hypomethylation of *F2RL3* in cancer cases relative to controls. This association is plausible as PAR-4 has been found to be involved in the process of blood coagulation as it is a thrombin receptor, and disruption of the normal coagulation process has been commonly described in lung cancer (94). Nonetheless, hypermethylation of one CpG site (chr19:17000567) within the *F2RL3* gene was found to be associated with an increased lung cancer risk. This result, while interesting, is difficult to explain. Hypermethylation of *F2RL3* has however been described in tissues of later stages of lung cancer, contrary to its hypomethylation observed in earlier stages tissues (54).

A priori, we hypothesized that hypomethylation of *AHRR* would result in its overexpression, which in turn, disrupts AHR transcriptional activity in the AhR pathway. It is thus expected that hypomethylation of *AHRR* would ensue an accumulation of environmental toxicants in the body which cannot be processed through the AhR pathway, leading to an increased lung cancer risk. We observed this association for the average measure of *AHRR* methylation as well as for 21 of 33 of the individual CpG sites in the *AHRR* gene. Nevertheless, in the individual CpG site analysis, we found that hypermethylation of one CpG site (chr5:369774) in the *AHRR* gene was associated with lung cancer risk. The contradicting association may be explained by the fact that *AHRR* has been suggested to play a role in the regulation of inflammatory responses (7). Through its interaction with various transcription factors such as NF-κB and HIF-1α, which are essential transcription factors in the regulation of inflammation, *AHRR* has been proposed to reduce apoptosis resistance, cell proliferation, and angiogenic and invasive growth by modulating the inflammation response to prevent the establishment of an auspicious environment for tumor

development (7,8). Hypermethylation of *AHRR*, and thus its downregulation, could therefore lead to an increased risk for cancer development due to the disruption of its regulation of the inflammatory response. While there are two possible mechanisms that could lead to carcinogenesis following aberrant methylation of *AHRR*, the association observed for the majority of the CpG sites in this study is indicative of hypomethylation among lung cancer cases compared to controls, and seems to support the first molecular mechanism described.

7.4 Study validity: methodological strengths and limitations

This section will cover the methodological strengths and limitations of our study with regard to its internal validity, while external validity will be covered later in the discussion.

7.4.1 Selection bias

A systematic difference between the characteristics of the participants selected for the study and of the population they are meant to represent is defined as selection bias. Traditional case-control studies are generally known to be susceptible to selection bias. The sampling of the participants in a typical prospective cohort study is independent of the outcome. Contrary to cohort studies, the challenge in case-control studies is to ensure that controls are sampled from the same source population of the cases.

Using a nested case-control design helps minimize selection bias. Participants are recruited from a defined study base, in this case, the CARTaGENE cohort, and thus, ensures that cases and controls arise from the same study base. Furthermore, given that CARTaGENE is a prospective cohort study, differential participation due to awareness of the outcome of interest is unlikely since the outcome of interest of the study is not yet known to the participants and researchers during enrolment. Selection bias was further minimized through random sampling of the controls for the study. Selection bias due to loss to follow-up is reduced as well since cases were identified through

linkage with the Quebec Cancer Registry and the RAMQ, a secure and valid administrative database of health information on Québec citizens (74). It is important to note that selection bias can also arise when participants are excluded based on criteria related to the exposure and the outcome, resulting in a systematic difference between groups. In our study, some participants were excluded due to missing methylation data. However, considering, the small and similar proportion of cases (4%) and controls (7%) excluded, and on the assumption that the data is missing at random (MAR) it is unlikely that it could have introduced a selection bias.

7.4.2 Measurement of outcome

The cases in our study were selected through linkage of CARTaGENE participants with the Quebec Cancer Registry (up to 2010) and the RAMQ, an administrative health database that uniformly covers the Québec population (74). However, RAMQ has recently been found to underestimate the true number of colorectal cases in the Québec population (95). While there is yet evidence of the RAMQ also underestimating the true number of lung cancer cases in the population, we should still consider its implication on the results of our study. It is possible that some undiagnosed cases were not included in our study. However, there is no reason to believe that being diagnosed (or not) in RAMQ as a case is related to exposure status. While we could also consider an alternate scenario whereby an undiagnosed case was selected as a control, this is unlikely as lung cancer is rare and symptoms are severe and thus, people are likely to seek care and thereby get a diagnosis.

7.4.3 Measurement of exposure

Information bias is a systematic bias that could emerge from measurement error or misclassification of the study variables. DNA methylation of *F2RL3* and *AHRR*, the exposure of interest, was measured using Sequenom EpiTYPER®, a method that combines bisulfite

sequencing and MALDI-TOF (matrix-assisted laser desorption/ionization coupled with time-of-flight) mass spectrometry. It is a validated and reliable high-throughput DNA methylation assay, which allows the quantification of DNA methylation at the single nucleotide level (75). Reliability of the DNA methylation measures was assessed by estimating their variability, and a CV of 4.65% between-plates and 4.16% between-fragments, based on the high-methylated DNA quality controls, lends confidence in the reliability of the measurements.

DNA methylation was measured in blood leukocytes, instead of directly in lung tissues. The use of whole blood samples as a surrogate for lung tissues permits the utilization of a relatively non-invasive, convenient, and inexpensive medium to quantify DNA methylation. While it is widely accepted that DNA methylation measurements can differ between cell types (96), multiple studies have shown that the use of blood samples to assess methylation is a reliable surrogate for DNA methylation in lung tissues (97–99). Considering that DNA methylation is a reversible process, its stability over time might be of concern. A study by Talens *et al.* compared DNA methylation at selected known loci involved in diseases in DNA samples that were collected 11 to 20 years apart. They found only minor DNA methylation differences between time points, indicating that DNA methylation is relatively stable over time (100). Given that DNA methylation is known to change with old age, it should be taken into account that those samples were taken in 34 participants aged between 14 and 64 years old from the Netherlands Twin Register biobank (100). The possibility of a greater change in DNA methylation over a longer period of time and older age cannot therefore be excluded. Another study that looked at global methylation in samples in an older population (60 to 87 years old at recruitment) did in fact observe a time-dependant change in DNA methylation between samples collected from visits separated on average by 11 to 16 years. Although, only 8-10% of the individuals in the study showed a methylation change >20%

(101). Both studies did however not address the possibility of DNA methylation changes due to developing diseases or changes in environmental and behavioural exposure in those individuals between time points. Taken together, these observations indicate that similar DNA methylation levels over time does not in itself guarantee DNA methylation stability as DNA methylation may change depending on the individual. Further investigations are therefore needed to examine DNA methylation stability over time with an attention to disease onset and exposure-related DNA methylation changes between time points.

The heterogeneous nature of whole blood specimens and the variability of blood cell type composition between individuals could introduce non differential classification due to measurement errors in the DNA methylation measurements as it is widely accepted that DNA methylation can differ between different cell types (99). Adjustment for cell type distribution was not possible in our study due to the absence of an external validation set to apply the algorithm for cell composition adjustment proposed by Houseman *et al*. (96). However, five EWASs which have looked at smoking exposures have noted that the variation in DNA methylation due to the heterogonous nature of whole blood is relatively small and insignificant when comparing methylation patterns between smokers and non-smokers (66,102–106). Furthermore, the results of our study are consistent with previous studies which have used the Houseman *et al*. algorithm to adjust for cell type composition, demonstrating that the impact of cell composition in whole blood samples on the associations observed in our study is likely minimal.

7.4.4 Implication of control selection strategy

Unlike a typical nested case-control study with risk set sampling, the controls in our study were sampled through cumulative sampling as they were sampled at the end of the follow-up. Indeed, risk set sampling selects controls who are at risk of developing the outcome at the same

time as the corresponding case, and who have not yet developed the outcome. However, as the controls were selected at the end of the cohort follow-up, our OR gives an estimate of the cohort OR, which would overestimates the strength of the association. However, since lung cancer is rare, the overestimation is minimal.

7.4.5 Confounding

Confounding refers to the "situation in which a noncausal association between a given exposure and an outcome is observed as a result of the influence of a third variable (or a group of variable), usually designated as a *confounding variable* or merely a *confounder*" (107).

The presence of confounders complicates the interpretation of the estimated association and poses a threat to the internal validity of a study. Our strategy to determine which confounding factors to include in our analysis consisted of a comprehensive literature review to create a DAG, a well-established knowledge-based causal diagram, to determine the minimal sufficient adjustment set. Based on the variables available in the CARTaGENE study, our adjusted set included age, sex, phase of blood sampling, fruit and vegetable consumption, BMI, ethnicity, and smoking. The possibility of residual confounding in our study is not excluded. It could have been introduced due to the re-categorization of multiple variables such as fruit and vegetable consumption, BMI, and ethnicity due to the small size of some categories. Residual confounding could have also occurred due to the unavailability in CARTaGENE and our study of data on confounding variables such as air pollution. Given that all participants in our study from similar environments (i.e. metropolitans), we expect this residual confounding to be negligible.

7.4.6 Temporality

A strength of this study is that it makes use of data collected prospectively, comparatively to traditional case-control studies. More precisely, this study used blood samples collected before

the diagnosis of lung cancer, ensuring the temporality between *F2RL3* and *AHRR* methylation, and lung cancer diagnosis.

7.5 External validity

External validity is defined as the extent to which the results of the study can be generalized to people and settings outside of the study population. CARTaGENE is the largest prospective health study in Quebec composed of men and women aged between 40 to 69 years old from six metropolitan areas of Quebec (Montreal, Quebec, Sherbrooke, Saguenay, Gatineau, and Trois-Rivières). The participants were randomly selected to be representative of the Quebec population based on provincial health registries, as such the characteristics of the cohort are representative of about 55.7% of the characteristics of the Quebec population (74). Considering the internal validity of our study due to the minimized selection bias, information bias, and confounding discussed above and that participants from the CARTaGENE cohort are largely representative of the Quebec population, it is reasonable to say that our study could be generalized to the Quebec population. This requires careful consideration given that DNA methylation is known to vary between ethnic groups (108,109), and that there are differences in the ethnical composition between our study population and Quebec's general population. Our study population mainly consisted of ethnically white participants (95%), which is slightly different from the proportion of Whites in Québec (84%), as of the 2021 Canadian census (110). Further analysis and stratifications could still be done to accurately extrapolate the results to Quebec's general population. However, the observed association that hypomethylation of *F2RL3* and *AHRR* increases lung cancer risk can safely be generalized to Quebec's population considering that the difference between the ethnical composition of the study's population and Quebec's general population is relatively small.

Similarly, comparable findings to our study are expected to be found within a population with the same ethnical composition. However, generalization to more ethnically diverse populations should be done with caution given that DNA methylation is known to differ by ethnicity (108,109). While we still expect an inverse association of *F2RL3* and *AHRR* methylation with lung cancer risk in more ethnically diverse population, it is possible that the magnitude of the association will diverge from the ones reported in this study due to the differing levels of exposure (i.e. DNA methylation) by ethnicities.

7.6 Conclusion and future directions

In this study, we showed further evidence that the methylation of *F2RL3* and *AHRR* could play a role in lung carcinogenesis. Given that smoking has been linked to the hypomethylation of *F2RL3* and *AHRR* and that it is likely to be upstream of the *F2RL3* and *AHRR* methylation-lung cancer association, future analysis should be carried out to clarify whether aberrant methylation in the two genes mediates the smoking-lung cancer association. Similarly, further investigations on the independent and combined impact of *F2RL3* and *AHRR* methylation with neighbouring genes could be of interest to explore other epigenetic gene-environment interactions in relation to lung carcinogenesis.

Our research adds to the existing body of knowledge regarding the association between intermediate endpoints and the risk of developing lung cancer. Such research on risk factors and underlying causal mechanisms is important to facilitate prevention initiatives as it can lead to the development of effective strategies for utilizing such markers in population health studies. Ultimately, enhanced knowledge regarding the fundamental mechanisms of lung cancer and its risk factors might lead to the development of more advanced screening tests in the future, and in time, a reduction of the burden of lung cancer on public health.

References

1. Wakelee HA, Chang ET, Gomez SL, Keegan TH, Feskanich D, Clarke CA, et al. Lung cancer incidence in never smokers. J Clin Oncol Off J Am Soc Clin Oncol. 2007;25:472–8.

2. Brzeziańska E, Dutkowska A, Antczak A. The significance of epigenetic alterations in lung carcinogenesis. Mol Biol Rep. 2013;40:309–25.

3. Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet. Nature Publishing Group; 2012;13:484–92.

4. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6.

5. Zhang Y, Yang R, Burwinkel B, Breitling LP, Holleczek B, Schöttker B, et al. F2RL3 methylation in blood DNA is a strong predictor of mortality. Int J Epidemiol. 2014;43:1215–25.

6. Gomides LF, Duarte ID, Ferreira RG, Perez AC, Francischi JN, Klein A. Proteinase-activated receptor-4 plays a major role in the recruitment of neutrophils induced by trypsin or carrageenan during pleurisy in mice. Pharmacology. 2012;89:275–82.

7. Vogel CFA, Haarmann-Stemmann T. The aryl hydrocarbon receptor repressor – More than a simple feedback inhibitor of AhR signaling: Clues for its role in inflammation and cancer. Curr Opin Toxicol. 2017;2:109–19.

8. Zudaire E, Cuesta N, Murty V, Woodson K, Adams L, Gonzalez N, et al. The aryl hydrocarbon receptor repressor is a putative tumor suppressor gene in multiple human cancers. J Clin Invest. 2008;118:640–50.

9. Mathers CD, Loncar D. Projections of Global Mortality and Burden of Disease from 2002 to 2030. PLOS Med. Public Library of Science; 2006;3:e442.

10. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. Int J Cancer. 2021;149:778–89.

11. Dela Cruz CS, Tanoue LT, Matthay RA. Lung Cancer: Epidemiology, Etiology, and Prevention. Clin Chest Med. 2011;32:10.1016/j.ccm.2011.09.001.

12. Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer. Canadian Cancer Statistics 2021. Toronto, ON: Statistics Canada and the Public Health Agency of Canada; 2021.

13. Canadian Cancer Statistics Advisory Committee in collaboration with the Canadian Cancer Society, Statistics Canada and the Public Health Agency of Canada. Canadian Cancer

Statistics: A 2020 special report on lung cancer [Internet]. Statistics Canada; 2020. Available from: https://www150.statcan.gc.ca/n1/daily-quotidien/200922/dq200922b-eng.htm

14. Barta JA, Powell CA, Wisnivesky JP. Global epidemiology of lung cancer. Ann Glob Health. 85.

15. Howlader N, Noone AM, Krapcho M, Neyman N, Aminou R, Waldron W, et al. SEER Cancer Statistics Review, 1975-2008. Bethesda, MD: National Cancer Institute; 2011.

16. Blandin Knight S, Crosbie PA, Balata H, Chudziak J, Hussell T, Dive C. Progress and prospects of early detection in lung cancer. Open Biol. Royal Society; 7:170070.

17. World Health Organization. WHO report on the global tobacco epidemic, 2011: warning about the dangers of tobacco. Geneva: World Health Organization; 2011.

18. Pesch B, Kendzia B, Gustavsson P, Jöckel K-H, Johnen G, Pohlabeln H, et al. Cigarette smoking and lung cancer – relative risk estimates for the major histological types from a pooled analysis of case-control studies. Int J Cancer. 2012;131:1210–9.

19. Tindle HA, Stevenson Duncan M, Greevy RA, Vasan RS, Kundu S, Massion PP, et al. Lifetime Smoking History and Risk of Lung Cancer: Results From the Framingham Heart Study. JNCI J Natl Cancer Inst. 2018;110:1201–7.

20. Risch HA, Howe GR, Jain M, Burch JD, Holowaty EJ, Miller AB. Are female smokers at higher risk for lung cancer than male smokers? A case-control analysis by histologic type. Am J Epidemiol. 1993;138:281–93.

21. Ragavan MV, Patel MI. Understanding sex disparities in lung cancer incidence: are women more at risk? Lung Cancer Manag. 9:LMT34.

22. North CM, Christiani DC. Women and Lung Cancer: What's New? Semin Thorac Cardiovasc Surg. 2013;25:10.1053/j.semtcvs.2013.05.002.

23. Wang M, Qin S, Zhang T, Song X, Zhang S. The effect of fruit and vegetable intake on the development of lung cancer: a meta-analysis of 32 publications and 20 414 cases. Eur J Clin Nutr. Nature Publishing Group; 2015;69:1184–92.

24. Wang C, Yang T, Guo X, Li D. The Associations of Fruit and Vegetable Intake with Lung Cancer Risk in Participants with Different Smoking Status: A Meta-Analysis of Prospective Cohort Studies. Nutrients. 2019;11:1791.

25. Zhang X, Liu Y, Shao H, Zheng X. Obesity Paradox in Lung Cancer Prognosis: Evolving Biological Insights and Clinical Implications. J Thorac Oncol. 2017;12:1478–88.

26. Duan P, Hu C, Quan C, Yi X, Zhou W, Yuan M, et al. Body mass index and risk of lung cancer: Systematic review and dose-response meta-analysis. Sci Rep. Nature Publishing Group; 2015;5:16938.

27. Yang Y, Dong J, Sun K, Zhao L, Zhao F, Wang L, et al. Obesity and incidence of lung cancer: A meta-analysis. Int J Cancer. 2013;132:1162–9.

28. Smith L, Brinton LA, Spitz MR, Lam TK, Park Y, Hollenbeck AR, et al. Body Mass Index and Risk of Lung Cancer Among Never, Former, and Current Smokers. JNCI J Natl Cancer Inst. 2012;104:778–89.

29. Schabath MB, Cress WD, Muñoz-Antonia T. Racial and Ethnic Differences in the Epidemiology of Lung Cancer and the Lung Cancer Genome. Cancer Control J Moffitt Cancer Cent. 2016;23:338–46.

30. Centers for Disease Control and Prevention (CDC). Racial/Ethnic disparities and geographic differences in lung cancer incidence --- 38 States and the District of Columbia, 1998-2006. MMWR Morb Mortal Wkly Rep. 2010;59:1434–8.

31. Campbell JD, Lathan C, Sholl L, Ducar M, Vega M, Sunkavalli A, et al. Comparison of Prevalence and Types of Mutations in Lung Cancers Among Black and White Populations. JAMA Oncol. 2017;3:801–9.

32. Mitchell KA, Zingone A, Toulabi L, Boeckelman J, Ryan BM. Comparative Transcriptome Profiling Reveals Coding and Noncoding RNA Differences in NSCLC from African Americans and European Americans. Clin Cancer Res Off J Am Assoc Cancer Res. 2017;23:7412–25.

33. Loomans-Kropp HA, Umar A. Cancer prevention and screening: the next step in the era of precision medicine. Npj Precis Oncol. Nature Publishing Group; 2019;3:1–8.

34. Rebbeck TR. Precision Prevention of Cancer. Cancer Epidemiol Biomarkers Prev. 2014;23:2713–5.

35. Meyskens FL, Mukhtar H, Rock CL, Cuzick J, Kensler TW, Yang CS, et al. Cancer Prevention: Obstacles, Challenges, and the Road Ahead. JNCI J Natl Cancer Inst. 2015;108:djv309.

36. Manser RL, Irving LB, Byrnes G, Abramson MJ, Stone CA, Campbell DA. Screening for lung cancer: a systematic review and meta-analysis of controlled trials. Thorax. BMJ Publishing Group Ltd; 2003;58:784–9.

37. National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, et al. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med. 2011;365:395–409.

38. Patz EF Jr, Pinsky P, Gatsonis C, Sicks JD, Kramer BS, Tammemägi MC, et al. Overdiagnosis in Low-Dose Computed Tomography Screening for Lung Cancer. JAMA Intern Med. 2014;174:269–74.

39. Anglim PP, Alonzo TA, Laird-Offringa IA. DNA methylation-based biomarkers for early detection of non-small cell lung cancer: an update. Mol Cancer. 2008;7:81.

40. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.

41. Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. Science. American Association for the Advancement of Science; 2017;

42. Brena RM, Costello JF. Genome–epigenome interactions in cancer. Hum Mol Genet. 2007;16:R96–105.

43. Baylin SB, Jones PA. Epigenetic Determinants of Cancer. Cold Spring Harb Perspect Biol. 2016;8.

44. Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nat Rev Genet. Nature Publishing Group; 2006;7:21–33.

45. Macaluso M, Paggi MG, Giordano A. Genetic and epigenetic alterations as hallmarks of the intricate road to cancer. Oncogene. Nature Publishing Group; 2003;22:6472–8.

46. Portela A, Esteller M. Epigenetic modifications and human disease. Nat Biotechnol. Nature Publishing Group; 2010;28:1057–68.

47. Kanai Y. Genome-wide DNA methylation profiles in precancerous conditions and cancers. Cancer Sci. 2010;101:36–45.

48. Kim YI, Giuliano A, Hatch KD, Schneider A, Nour MA, Dallal GE, et al. Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. Cancer. 1994;74:893–9.

49. Towle R, Truong D, Hogg K, Robinson WP, Poh CF, Garnis C. Global analysis of DNA methylation changes during progression of oral cancer. Oral Oncol. 2013;49:1033–42.

50. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. Future Medicine; 2009;1:239–59.

51. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. Nature Publishing Group; 2002;3:415–28.

52. National Library of Medicine (US). Homo sapiens F2R like thrombin or trypsin receptor 3 (F2RL3), RefSeqGene (LRG_633) on chromosome 19 [Internet]. GenBank Natl. Cent. Biotechnol. Inf. NCBI. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NG_032894.1?from=4985&to=8592&report=genbank

53. Xu WF, Andersen H, Whitmore TE, Presnell SR, Yee DP, Ching A, et al. Cloning and characterization of human protease-activated receptor 4. Proc Natl Acad Sci U S A. 1998;95:6642–6.

54. Jiang P, Yu G-Y, Zhang Y, Xiang Y, Hua H-R, Bian L, et al. Down-regulation of protease-activated receptor 4 in lung adenocarcinoma is associated with a more aggressive phenotype. Asian Pac J Cancer Prev APJCP. 2013;14:3793–8.

55. Ghio P, Cappia S, Selvaggi G, Novello S, Lausi P, Zecchina G, et al. Prognostic role of protease-activated receptors 1 and 4 in resected stage IB non-small-cell lung cancer. Clin Lung Cancer. 2006;7:395–400.

56. Barradas M, Monjas A, Diaz-Meco MT, Serrano M, Moscat J. The downregulation of the pro-apoptotic protein Par-4 is critical for Ras-induced survival and tumor progression. EMBO J. 1999;18:6362–9.

57. Diaz-Meco MT, Moscat J. Akt regulation and lung cancer: A novel role and mechanism of action for the tumor suppressor Par-4. Cell Cycle. Taylor & Francis; 2008;7:2817–20.

58. National Library of Medicine (US). Homo sapiens aryl-hydrocarbon receptor repressor (AHRR), RefSeqGene on chromosome 5 [Internet]. GenBank Natl. Cent. Biotechnol. Inf. NCBI. Available from: https://www.ncbi.nlm.nih.gov/nuccore/NG_029834.2?from=5001&to=139116&report=genbank

59. Hahn ME, Allan LA, Sherr DH. Regulation of Constitutive and Inducible AHR Signaling: Complex Interactions Involving the AHR Repressor. Biochem Pharmacol. 2009;77:485–97.

60. Larigot L, Juricek L, Dairou J, Coumoul X. AhR signaling pathways and regulatory functions. Biochim Open. 2018;7:1–9.

61. Murray IA, Patterson AD, Perdew GH. Aryl hydrocarbon receptor ligands in cancer: friend and foe. Nat Rev Cancer. Nature Publishing Group; 2014;14:801–14.

62. Baglietto L, Ponzi E, Haycock P, Hodge A, Bianca Assumma M, Jung C, et al. DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk. Int J Cancer. 2017;140:50–61.

63. Schlosberg CE, VanderKraats ND, Edwards JR. Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Nucleic Acids Res. 2017;45:5100–11.

64. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011;88:450–7.

65. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS One. 2013;8:e63812.

66. Guida F, Sandanger TM, Castagné 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.

67. Zhang Yan, Yang Rongxi, Burwinkel Barbara, Breitling Lutz P., Brenner Hermann. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. Environ Health Perspect. Environmental Health Perspectives; 2014;122:131–7.

68. Alhamdow A, Lindh C, Hagberg J, Graff P, Westberg H, Krais AM, et al. DNA methylation of the cancer-related genes F2RL3 and AHRR is associated with occupational exposure to polycyclic aromatic hydrocarbons. Carcinogenesis. 2018;39:869–78.

69. Tantoh DM, Wu M-C, Chuang C-C, Chen P-H, Tyan YS, Nfor ON, et al. AHRR cg05575921 methylation in relation to smoking and PM2.5 exposure among Taiwanese men and women. Clin Epigenetics. 2020;12:117.

70. Arechederra M, Daian F, Yim A, Bazai SK, Richelme S, Dono R, et al. Hypermethylation of gene body CpG islands predicts high dosage of functional oncogenes in liver cancer. Nat Commun. Nature Publishing Group; 2018;9:3164.

71. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. 2009;41:178–86.

72. Rao X, Evans J, Chae H, Pilrose J, Kim S, Yan P, et al. CpG island shore methylation regulates caveolin-1 expression in breast cancer. Oncogene. Nature Publishing Group; 2013;32:4519–28.

73. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene Body Methylation can alter Gene Expression and is a Therapeutic Target in Cancer. Cancer Cell. 2014;26:577–90.

74. Awadalla P, Boileau C, Payette Y, Idaghdour Y, Goulet J-P, Knoppers B, et al. Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. Int J Epidemiol. Oxford Academic; 2013;42:1285–99.

75. Suchiman HED, Slieker RC, Kremer D, Slagboom PE, Heijmans BT, Tobi EW. Design, measurement and processing of region-specific DNA methylation assays: the mass spectrometry-based method EpiTYPER. Front Genet. 2015;6.

76. Ho V, Ashbury JE, Taylor S, Vanner S, King WD. Quantification of gene-specific methylation of DNMT3B and MTHFR using sequenom EpiTYPER®. Data Brief. 2016;6:39–46.

77. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. American Association for the Advancement of Science; 2012;337:1190–5.

78. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. Nature Publishing Group; 2011;480:490–5.

79. Kyung Lee M, Armstrong DA, Hazlett HF, Dessaint JA, Mellinger DL, Aridgides DanielS, et al. Exposure to extracellular vesicles from Pseudomonas aeruginosa result in loss of DNA methylation at enhancer and DNase hypersensitive site regions in lung macrophages. Epigenetics. 16:1187–200.

80. Menard S. Standards for Standardized Logistic Regression Coefficients. Soc Forces. 2011;89:1409–28.

81. Vieira AR, Abar L, Vingeliene S, Chan DSM, Aune D, Navarro-Rosenblatt D, et al. Fruits, vegetables and lung cancer risk: a systematic review and meta-analysis. Ann Oncol Off J Eur Soc Med Oncol. 2016;27:81–96.

82. Ghazi T, Arumugam T, Foolchand A, Chuturgoon AA. The Impact of Natural Dietary Compounds and Food-Borne Mycotoxins on DNA Methylation and Cancer. Cells. 2020;9:2004.

83. Hoffmann K, Krause C, Seifert B. The German Environmental Survey 1990/92 (GerES II): primary predictors of blood cadmium levels in adults. Arch Environ Health. 2001;56:374–9.

84. Greenland, Sander. Analysis of polytomous exposures and outcomes, in: K.J.Rothman, S. Greenland (Eds.). Mod Epidemiol. 2nd ed. Philadelphia: Lippincott Williams &Wilkins; 1998. page 301.

85. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. [Royal Statistical Society, Wiley]; 1995;57:289–300.

86. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. Behav Brain Res. 2001;125:279–84.

87. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans R Soc Math Phys Eng Sci. Royal Society; 2016;374:20150202.

88. Battram T, Richmond RC, Baglietto L, Haycock PC, Perduca V, Bojesen SE, et al. Appraising the causal relevance of DNA methylation for risk of lung cancer. Int J Epidemiol. 2019;48:1493–504.

89. Zhang Y, Schöttker B, Ordóñez-Mena J, Holleczek B, Yang R, Burwinkel B, et al. F2RL3 methylation, lung cancer incidence and mortality. Int J Cancer. 2015;137:1739–48.

90. Davey Smith G, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. Int J Epidemiol. 2003;32:1–22.

91. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012;41:161–76.

92. Grieshober L, Graw S, Barnett MJ, Thornquist MD, Goodman GE, Chen C, et al. AHRR methylation in heavy smokers: associations with smoking, lung cancer risk, and lung cancer mortality. BMC Cancer. 2020;20:905.

93. Jacobsen KK, Schnohr P, Jensen GB, Bojesen SE. AHRR (cg05575921) Methylation Safely Improves Specificity of Lung Cancer Screening Eligibility Criteria: A Cohort Study. Cancer Epidemiol Biomarkers Prev. 2022;31:758–65.

94. Ferrigno D, Buccheri G, Ricca I. Prognostic significance of blood coagulation tests in lung cancer. Eur Respir J. 2001;17:667–73.

95. Diop M, Strumpf EC, Datta GD. Measuring colorectal cancer incidence: the performance of an algorithm using administrative health data. BMC Med Res Methodol. 2018;18:38.

96. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.

97. Stueve TR, Li W-Q, Shi J, Marconett CN, Zhang T, Yang C, et al. Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Hum Mol Genet. 2017;26:3014–27.

98. de Vries M, van der Plaat DA, Nedeljkovic I, Verkaik-Schakel RN, Kooistra W, Amin N, et al. From blood to lung tissue: effect of cigarette smoke on DNA methylation and lung function. Respir Res. 2018;19:212.

99. Houseman EA, Kim S, Kelsey KT, Wiencke JK. DNA Methylation in Whole Blood: Uses and Challenges. Curr Environ Health Rep. 2015;2:145–54.

100. Talens RP, Boomsma DI, Tobi EW, Kremer D, Jukema JW, Willemsen G, et al. Variation, patterns, and temporal stability of DNA methylation: considerations for epigenetic epidemiology. FASEB J Off Publ Fed Am Soc Exp Biol. 2010;24:3135–44.

101. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, et al. Intra-individual change in DNA methylation over time with familial clustering. JAMA J Am Med Assoc. 2008;299:2877–83.

102. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics. 2015;7.

103. Tsaprouni LG, Yang T-P, Bell J, Dick KJ, Kanoni S, Nisbet J, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9:1382–96.

104. Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans. Hum Genet. 2013;132:1027–37.

105. Besingi W, Johansson Å. Smoke-related DNA methylation changes in the etiology of human disease. Hum Mol Genet. 2014;23:2290–7.

106. Dogan MV, Shields B, Cutrona C, Gao L, Gibbons FX, Simons R, et al. The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women. BMC Genomics. 2014;15:151.

107. Szklo M (Moyses), Nieto FJ. Epidemiology : beyond the basics. Fourth edition. Burlington, Massachusetts: Jones & Bartlett Learning; 2019.

108. Zhang FF, Cardarelli R, Carroll J, Fulda KG, Kaur M, Gonzalez K, et al. Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood. Epigenetics. 2011;6:623–9.

109. Park SL, Patel YM, Loo LWM, Mullen DJ, Offringa IA, Maunakea A, et al. Association of internal smoking dose with blood DNA methylation in three racial/ethnic populations. Clin Epigenetics. 2018;10:110.

110. Statistics Canada. Profile table, Census Profile, 2021 Census of Population - Quebec [Province] [Internet]. 2022 Feb. Available from: https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/prof/index.cfm?Lang=E