Applications of Genomic and Epigenomic Signatures to Identify Markers of Exogenous Exposures and Elucidate their Potential Role in Cancer Aetiology

Hanane Omichessan

université PARIS-SACLAY

NNT : 2019SACLS558

Thèse de doctorat

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir

Applications of genomic and epigenomic signatures to identify markers of exogeneous exposures and elucidate their potential role in cancer aetiology

University Paris-Saclay Doctoral Thesis
Prepared at the University of Paris-Sud

Doctoral school n°570 EDSP | Public Health
Doctoral Specialisation: Biostatistics

Thesis presented and defended in Villejuif, the 17th of December 2019, by

Hanane Omichessan

Jury composition:

Paolo Provero Professor, University of Torino (Torino, Italy)	**President and Reporter**
Marie-Aline Charles Director of research, INSERM (Paris, France)	**Reporter**
Valérie Chaudru Associate professor, University of Evry (Evry, France)	**Examiner**
Johanna Lepeule Researcher, University of Grenoble (Grenoble, France)	**Examiner**
Gianluca Severi Director of recherche, INSERM (Villejuif, France)	**Thesis director**
Vittorio Perduca Associate professor,University of Paris-Descartes (Paris, France)	**Thesis co-director**

Thesis prepared in the framework of the Public Health Doctoral Network coordinated by École des Hautes Études en Santé Publique (EHESP) and prepared within « Health across generations » team of the Center for Research in Epidemiology and Population Health (CESP), INSERM U1018:

Gustave Roussy – Espace Maurice Tubiana
114 rue Edouard Vaillant
94 805 Villejuif cedex

FOREWORD

Following studies in biology carried out in Benin, then in molecular biology at the University of Evry Val d'Essonne, I started my training in bioinformatics and associated fields in 2014 by integrating the Master 1 mention bioinformatics, GENomics, Informatics and Mathematics for Health and Environment (GENIOMHE) of Paris-Saclay University.

During this course, I realized two internships including one in Germany at the department of bioinformatics within the Institute for Microbiology and Genetics, a component of Georges-August University of Göttingen. The second internship was with the INSERM U1018, "Health across generations" team of Gustave Roussy Institute, directed by Dr. Gianluca Severi. Under his supervision, I investigated the association between circulating levels of B vitamins and DNA methylation.

The "Health across generations" team conducts research projects related to the identification and analysis of the role of environment and lifestyle in the occurrence of women's cancers and other non-communicable diseases through E3N, a prospective cohort of almost 100.000 women. The team has recently started the recruitment of their husbands (E4N-G1), children (E4N-G2) and grandchildren (E4N-G3).

My pre-doctoral internships allowed me to gain experience in the analysis of genomics, epigenomics and epidemiological data and in the design of related studies. Following the obtention in july 2016 of a grant from the French National Institute of Cancer (INCa), I wanted to continue my research in the "Health across generations" team.

I did my thesis under the joint supervision of Drs. Gianluca Severi and Vittorio Perduca.

My doctoral work has been focused on the applications of genomic and epigenomic signatures to identify markers of exogenous exposures and elucidate their potential role in cancer aetiology. Data used included simulations, public repositories such as The Cancer Genome Atlas and those from to the French E3N prospective cohort.

This thesis is divided into 5 chapters. After a review of the concepts related to my work, recent advances in the study of mutational and epigenetic signatures in tumours will be described, followed by a chapter covering one most the most recent developments with regards to cancer genomics. The fourth chapter will report the investigations performed for the identification of novel markers of exposition to endocrine disruptors. And finally, a summary of the findings and the research perspectives will be presented.

ABSTRACT

Background: Several risks factors have been identified for cancer, and it has been estimated that more than 40% of cases in developed countries are preventable through the modulation of known modifiable risk factors.

Objectives: The overall objective of this thesis was to demonstrate that the analysis of genomic and epigenomic data integrated with well-characterised exposure and lifestyle data may be used to identify markers of environmental exposures and lifestyle and may contribute to increase our understanding of cancer aetiology.

Results: We first describe how genomic and epigenomic signatures can be used to identify markers of exposure and decipher the aetiology of cancer. Then, we adopt the mutational signatures framework to contribute to the debate about the "bad luck" hypothesis for cancer and demonstrate that tobacco-related mutations are more strongly correlated with cancer risk than random mutations. We introduce a probabilistic model for the simulation of mutational signature data and compare the performance of the available methods for the identification of mutational signatures using both simulated and real data. Additionally, we introduce a new method for the identification of such signatures. Finally, we use methylation array data in an epidemiological study within the E3N cohort to investigate the association between exposure to Brominated Flame Retardants and Per- and polyfluoroalkyl substances, two organic pollutants that are known endocrine disrupting chemicals, and methylation in DNA from blood. Overall, our study does not provide evidence of methylation alterations at the level of the whole genome, in regions or in single CpGs. Suggestive evidence of alterations in the methylation of genes within plausible biological pathways (e.g. androgen response) warrants further investigations.

Conclusions: Our work on the methodological aspects of mutational signature research introduces an original framework for measuring the performance of tools for the identification of mutational signatures that may serve as reference for future methodological or applied research. Our applications of both mutational signature and methylome research demonstrate the usefulness of such tools to assess exposures and elucidate their role in cancer aetiology.

Keywords : mutational signatures, DNA methylation, endocrine disruptors, epidemiology, lifestyle

RESUME

Contexte : Plusieurs facteurs de risque de cancer ont été identifiés et il a été estimé que plus de 40% des cas dans les pays développés pourraient être évités en modifiant les facteurs de risque connus

Objectifs : L'objectif général de cette thèse était de démontrer que l'intégration de données génomiques et épigénomiques aux données détaillées sur les expositions environnementales et le mode de vie peut être utile pour identifier des biomarqueurs de ces facteurs et contribuer à augmenter notre connaissance de l'étiologie du cancer.

Résultats : Dans un premier temps, nous décrivons comment les signatures génomiques et épigénomiques peuvent être utilisées pour identifier des marqueurs d'exposition et déchiffrer l'étiologie du cancer. Ensuite, nous contribuons au débat relatif à l'hypothèse de la chance dans le développement du cancer et démontrons que les mutations induites par le tabagisme sont plus prédictives du risque de cancer que les mutations aléatoires. Nous introduisons un modèle probabiliste pour la simulation de données mutationnelles et comparons la performance des outils d'identification de ces signatures avec des données réelles et simulées. De plus, nous introduisons une nouvelle méthode pour l'identification des signatures mutationnelles. Enfin, nous utilisons les données de méthylation de la cohorte E3N pour étudier le lien entre l'exposition aux retardateurs de flamme bromés et aux composés perfluorés, deux substances classées parmi les perturbateurs endocriniens, et la méthylation de l'ADN sanguin. Globalement, notre étude ne fournit aucune preuve d'altérations globales du méthylome ou d'altérations à l'échelle des CpGs. Cependant, certains résultats suggèrent l'existence d'altérations de la méthylation de gènes impliqués dans des voies biologiques (ex., la réponse aux androgènes) et nécessitent des recherches supplémentaires**.**

Conclusions : Ce travail contribue à la recherche méthodologique portant sur les signatures mutationnelles en introduisant un protocole de mesure de performance et d'identification des signatures mutationnelles pouvant servir de référence à de futures études méthodologiques ou appliquées. Nos recherches sur les signatures mutationnelles et le méthylome démontrent l'utilité de tels outils pour évaluer les expositions et élucider leur rôle dans l'étiologie du cancer.

Mots clés : signatures mutationnelles, méthylation de l'ADN, perturbateurs endocriniens, épidémiologie, mode de vie

AKNOWLEDGEMENTS

Foremost, I would like to extend my deepest thanks to my two supervisors, Dr. Gianluca Severi and Dr. Vittorio Perduca, for their involvement, enthusiasm, and constant encouragement and support at each phase of this PhD. I sincerely thank both of you for giving me the opportunity to work with you, for your precious research ideas and for all that I learnt during these years.

I sincerely thank Prof. Marine-Aline Charles for accepting to be part of my jury panel and to review my PhD thesis and Drs. Valéry Chaudru and Johanna Lepeule for accepting the role of examiners. I am also very grateful to Prof. Paolo Provero which in addition to being *rapporteur*, accepted the role of the president of my thesis committee. I am very honoured to have my PhD reviewed by such experienced researchers.

Thanks also to the Reviewers of the papers that have been published out of this PhD for their high-quality contribution that helped to improve greatly these articles, and thereby this thesis.

I wish to acknowledge the *Institut National du Cancer* (INCa) and the E3N team for their financial support during my PhD, as well as well as *École des Hautes Études en Santé Publique* (EHESP) for the funding of my travels within France, and between Paris, UK and Greece. My sincere thanks go to *École Doctorale de Santé Publique* (EDSP) of Paris-Saclay University, its former director Prof. Jean Bouyer, the current Prof. Florence Ménégaux and Fabienne Renoirt for their availability and support in the administrative aspects.

I would like to thank Dr. Marie-Christine Boutron-Ruault, the former director of the Health across Generations team, for welcoming me, first for a master internship, and finally during my PhD. I also sincerely thank Drs. Laura Baglietto and Francesca Manicini, for their valuable involvement, suggestions and recommendations. Thank you, Dr. Fanny Artaud, for your precious advices and Dr. Marina Kvaskoff for your kindness and friendship, and for welcoming me in your mentoring program.

I would like to express my gratitude to Drs. Tania Di Gioia and Jessica Pericaud for welcoming me within their team and giving me the opportunities to have an overview of entrepreneurship, innovation and valorisation.

Thanks to my mentors Dr. Grégory Peignon and Françoise Touboul for their encouragement and support during this project and for the future prospects.

I thank all members of E3N team, for their help, support, and encouragement. I couldn't have asked for better colleagues! A particular thank to Iris, my PhD twin who started this adventure at the same time with me three years ago. Thank you for all those crazy moments, these ups and downs shared together. Thank you, Roselyn, Sofiane, Doua, Mahamat, Emmanuelle, Amandine, Marie and Nasser for all the moments we shared together. I am truly grateful for your availability, kindness and friendship.

A special thanks to Solène, Emeline, Armelle, Monia, Fatou and Charlotte. It was amazing to work and have fun with you and others PhD candidate of EDSP, but also *for Confédération des Jeunes Chercheurs* (CJC). I really enjoyed these dinner, bowling, party and among others discussions about PhD'careers.

Thank you, Imane and Amira, for these laughs, smile, jokes. Thanks to my childhood friends and to all the others who have, from near and far, never stopped supporting me.

And of course, thank you, TH, for your constant support and patience.

My ultimate thanks go to my parents, brother and sisters for their love and patience, particularly during the last phase of this PhD.

Thank you, mum, for your unconditional support, which have been essential all the way through.

I am grateful to all those who will be interested in this doctoral work and who will read part or all of this manuscript.

SCIENTIFIC PRODUCTION

PUBLISHED WORK

Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumors: a review and empirical comparison of performance. PLoS One. 2019 Sep 12;14(9): e0221235

Perduca V, **Omichessan H**, Baglietto L, Severi G. Mutational and epigenetic signatures in cancer tissue linked to environmental exposures and lifestyle. *Curr Opin Oncol. 2018 Jan;30(1):61-67*

Perduca V, Alexandrov LB, Kelly-Irving M, Delpierre C, **Omichessan H**, Little MP, Vineis P, Severi G. Stem cell replication, somatic mutations and role of randomness in the development of cancer. *Eur J Epidemiol. 2019 Jan 8*

ARTICLES SUBMITTED

Mancini FR, Cano-Sancho G, Mohamed O, Cervenka I, **Omichessan H,** Marchand P, Boutron-Ruault MC, Arveux P, Severi G, Antignac JP, Kvaskoff M. Plasma concentration of brominated flame retardants and breast cancer risk: A nested case-control study in the French E3N cohort.

ARTICLES WITH SUBMISSION IN PROGRESS

Omichessan H, Perduca V, Mancini FR, Baglietto L, Severi G. Association between Brominated Flame Retardants and DNA methylation

Omichessan H, Perduca V, Mancini FR, Baglietto L, Severi G. Association between Per- and polyfluorinated Alkylated Substances and DNA methylation

OTHER PUBLICATIONS

Fedirko V, Jenab M, Méplan C, Jones JS, Zhu W, Schomburg L, Siddiq A, Hybsier S, Overvad K, Tjønneland A, **Omichessan H** et al.. Association of selenoprotein and selenium pathway genotypes with risk of colorectal cancer and interaction with Selenium status. *Nutrients. 2019 Apr 25;11(4). pii: E935*

Schmit SL, Edlund CK, Schumacher FR, Gong J, …, **Omichessan H** et al. Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *J Natl Cancer Inst. 2019 Feb 1;111(2):146-157. doi: 10.1093/jnci/djy099. PubMed PMID: 29917119*

Campa D, Barrdahl M, Santoro A, Severi G, Baglietto L, **Omichessan H** et al. Mitochondrial DNA copy number variation, leukocyte telomere length, and breast cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC) study. *Breast Cancer Res. 2018 Apr 17;20(1):29*

Perrier F, Novoloaca A, Ambatipudi S, Baglietto L, Ghantous A, Perduca V, Barrdahl M, Harlid S, Ong KK, Cardona A, Polidoro S, Nøst TH, Overvad K,**Omichessan H** et al. Identifying and correcting epigenetics measurements for systematic sources of variation. *Clin Epigenetics. 2018 Mar 21;10:38.*

COMMUNICATIONS

Omichessan H, Severi G, Perduca V. Deciphering the signatures of mutational process in human cancer: a review of algorithms and methods Rencontres scientifiques de l'EHESP – ISPED (13-14/03/18)

Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance – EACR 4th Conference on Cancer Genomics. Churchill College - Cambridge (23-26/06/19)

Omichessan H, Severi G, Perduca V. Computational tools to detect signatures of mutational processes in DNA from tumours: a review and empirical comparison of performance – Mutographs International meeting. IARC - Lyon (11-12/07/19)

CO-SUPERVISION

Gabriele Alaimo. Master internship in molecular biotechnology – University of Torino
Mutational signatures in cancer genomes and application to mesothelioma

ACTIVITIES OUTSIDE RESEARCH

DOCTORAL MISSION

Prospection officer
Direction de l'Orientation Professionnelle et des Relations Entreprises Paris-Sud (32 days)
• Marketing of training offer, cartography of training and associated skills
• Organization of events to promote professional insertion of young graduates

Valorisation officer
Société d'Accélération du Transfert de Technologies Paris-Saclay (32 days)
• Presentation of the innovation offer within Paris-Saclay laboratories
• Promotion of the web portal with corporates and valorization at innovation show such as TechInnov or Vivatechnogy

VULGARISATION AND PROMOTION

Co-organizer of events related among others to health prevention and PhD careers.

ASSSOCIATION

Association des doctorants et docteurs de l'École Doctorale de Santé Publique : *in charge of extra-scientific activities, then vice-president since April 2018*

TABLE OF CONTENTS

Foreword ... 1
Abstract ... 3
Résumé ... 5
Aknowledgements ... 7
Scientific production ... 9
Activities outside research ... 11
Table of Contents ... 13
List of Figures ... 19
List of Tables ... 21
List of Appendices ... 23
List of Abbreviations ... 25
Chapter I: ... 27
General introduction ... 27
1. Genomic signatures ... 31
1.1 Behind the concept of "mutational signatures" ... 31
1.1.1 Hallmarks of cancer ... 31
1.1.2 Somatic mutations and related therories ... 32
1.1.3 Base substitutions and genomic alterations ... 33
1.2 Mathematical modeling of a mutational process ... 35
1.2.1 Definition of mutational catalogues, spectra and signatures ... 35
1.2.2 Deciphering the signatures of mutational processes: *de novo vs. refitting* ... 37
1.3 COSMIC: Catalogue Of Somatic Mutations In Cancer ... 40
1.4 Experimental validation of mutational signatures ... 43
2. Epigenomic signatures ... 44
2.1 Introduction to epigenetics ... 44
2.1.1 Overview ... 44
2.1.2 DNA methylation and epigenetic mechanisms ... 45
2.2 Profiling DNA methylation ... 48
2.2.1 Methodological aspects ... 48
2.2.2 Beta-values and M-values in microarray analysis ... 49
2.3 How does lifestyle influence DNA methylation ... 51
3. Endocrine disruptors ... 52
3.1 Introduction to Persistent Organic Pollutants ... 52
3.1.1 Brominated Flame Retardants (BFRs) ... 53
3.1.2 Per- and polyfluorinated Alkylated Substances (PFASs) ... 58

3.2 Persistent Organic Pollutants and DNA methylation 62
4. Summary and Objectives 64
Chapter II: 67
Environment and lifestyle influence on molecular features 67
1. Environmental exposures associated mutational and epignetics signatures 71
1.1The exogeneous causes of mutational signatures 71
1.1.1 Tobacco 71
1.1.2 Aflatoxin B1 72
1.1.3 Ionizing radiation 72
1.1.4 UV light 72
1.1.5 Aristolochic acid 73
1.2 Exposures related epigenetics signatures in tumour tissue 74
2. Exposure to smoking, lung adenocarcinoma development and the "bad" luck cancer theory 76
2.1 The "bad luck" debate: stem cell divisions, driver mutations and cancer risk 77
2.2 Predicting lung cancer risk via *extrinsinc* mutations 80
3. Conclusion 82
Chapter III: 83
Computational tools to detect signatures of mutational process 83
1. Context 87
2. Overview of avalaible tools for mutational signature analysis 88
2.1 *De novo* approaches 92
2.2 Refitting with known mutational signatures 93
2.3 Combining de novo and refitting procedure 94
3. Materials and experimental settings 95
3.1 The Cancer Genome Atlas 95
3.2 Our original refitting tool: MutationalCone 95
3.3 Simulation of a mutational catalogue 96
4. Comparison of algorithms performance 100
4.1. Specificity and sensitivity for *de novo* extraction and assignment 100
4.2 Bias of refitting procedures 101
5. Findings 102
5.1 Performance of *de novo* tools 102
5.1.1 Frobenius norm 102
5.1.2 Confusion matrices 104
5.2 Performance of refitting tools 110
6. Conclusion 113

Chapter IV: **115**
Association between Persistent Organic Pollutants and DNA methylation **115**
1. Materials: the E3N prospective cohort **119**
1.1 Presentation of the cohort 119
1.2 Epidemiological data collected in E3N 119
1.2.1 Data collection 119
1.2.2 Dietary questionnaire 122
1.2.3 The E3N-TDS2 database on individual exposure to contaminants 122
1.3 Measurement of circulating levels of BFRs and PFASs 124
1.3.1 Design of the case-control study 124
1.3.2 Circulating levels of BFRs 124
1.3.3 Circulating levels of PFASs 125
1.4 Assessing DNA methylation in E3N 125
2. Statistical analyses **127**
2.1. Descriptive statistics 127
2.1.1 Median, frequency and other basics statistics 127
2.1.2 Quantile-quantile plot 127
2.2 Association measures 127
2.2.1 Fixed vs. random effects 128
2.2.2 Mathematical definition of a linear mixed effects models 129
2.2.3 Statistical modeling 130
2.2.4 False Discovery Rate 130
2.2.5 Missing data 131
2.3 Gene Set Enrichment Analysis 131
2.3.1 Overview 131
2.3.2 The Molecular Signature Database 132
3. Methylation signatures of Brominated Flame Retardants **133**
3.1 Approaches 133
3.1.1 Association between dietary exposure to BFRs and DNA methylation 133
3.1.2 Association between circulating levels of BFRs and DNA methylation 133
3.1.3 Enrichment analysis 134
3.2 Findings 134
3.2.1 Baseline characteristics of the study population 134
3.2.2 Epigenome-wide association study: BFRs and methylation of blood DNA 137
3.2.3 BFRs and global or regional methylation 140
3.2.4 BFRs and methylation alteration in specific pathways: Gene Set Enrichment Analyses 144
4. Methylation signatures of Per- and polyfluorinated Alkylated Substances **146**
4.1 Approaches 146

4.1.1 Association between dietary exposure to PFASs and DNA methylation 146
4.1.2 Association between circulating levels of PFASs and DNA methylation 146
4.1.3 Enrichment analysis 147
4.2 Findings 147
4.2.1 Baseline characteristics of the study population 147
4.2.2 Epigenome-wide association study: PFASs and methylation of blood DNA 149
4.2.3 PFASs and global or regional methylation 150
4.2.4 PFASs and methylation alterations in specific pathways: Gene Set Enrichment Analysis 154
5. Conclusion 155
5.1 Methylation signatures of Brominated Flame Retardants 155
5.2 Methylation signatures of Per- and polyfuorinated alkylated substances 156
Chapter V: 159
General discussion and future prospects 159
1. Synthesis 161
1.1 Genomic signatures 161
1.2 Epigenomic signatures 162
2. Research perspectives 163
2.1 Genomic signatures 163
2.2 Epigenomic signatures 163
3. Implication in public health 165
Appendices 167
1. Introduction 171
1.1 Les signatures mutationnelles 171
1.2 La méthylation de l'ADN 172
1.3 Les polluants organiques persistants 172
1.4 Objectifs 172
2. Matériels et méthodes 174
2.1 Identification des signatures mutationnelles 174
2.1.1 Aperçu des méthodes existantes 174
2.1.2 Simulation d'un catalogue mutationnel 174
2.1.3 La base de données TCGA 175
2.1.4 Évaluation de la performance des méthodes 175
2.2 Association entre pertubateurs endocriniens et méthylation de l'ADN 175
2.2.1 La cohorte E3N 175
2.2.2 Collection des données 175
2.2.3 Mesure du niveau circulants des BFRs et des PFASs 176
2.2.4 Méthylation de l'ADN 176

2.2.5 Gene Set Enrichment Analysis 176
2.2.6 Analyses statistiques 176
3. Résultats 177
3.1 Expositions environmentales associées aux signatures moléculaires 177
3.1.1 Expositions environmentales associées aux signatures mutationnelles et épigénétiques 177
3.1.2 Tabagisme, cancer du poumon et la role de la chance dans le dévelopement du cancer 177
3.2 Performance des algorithmes d'identification des signatures mutationnelles 178
3.3 Association entre pertubateurs endocriniens et méthylation de l'ADN 179
3.3.1 Association entre BFRs et méthylation de l'ADN 179
3.3.2 Association entre PFASs et méthylation de l'ADN 179
4. Discussion et conclusion 180
4.1 Expositions environmentales associées aux signatures moléculaires 180
4.2 Performance des algorithmes d'identification des signatures mutationnelles 180
4.3 Association entre pertubateurs endocriniens et méthylation de l'ADN 181
References 201

LIST OF FIGURES

Figure I.1. The transformation process of normal cells to malignant cells. 31
Figure I.2. Somatic mutations leading to carcinogenesis 32
Figure I.3. 100 years of somatic mutations theory 33
Figure I.4. The 96 mutations types in a trinucleotide context 34
Considerations of the 6 types of base substitutions_ a DNA base is replaced by another (C>A, C>G, C>T, T>A, T>C and T>G) and the associated sequence context. 34
Figure I.5. Mutational catalogue and the individual signatures contribution to it 36
Figure I.6. Comparison of newly identified signatures with COSMIC signatures 38
Figure I.7. Cosine similarity plot of COSMIC signatures 39
Figure I.8. Overview of COSMIC tools 40
Figure I.9. Patterns of mutational signatures (v2 – March 2015): 30 SBS 41
Figure I.10. Patterns of mutational signatures (v3 – May 2019) : 49 SBS 42
Figure I.11. DNA methylation 45
Figure I.12. Micronutrient donors involved in one-carbon metabolism and subsequently in DNA methylation (one-carbon metabolism) 46
Figure I.13. Effect of DNA methylation on gene expression 47
Figure I.14. Evolution of next-generation sequencing-based techniques applied to DNA methylation profiling. 48
Figure I.15. Main DNA methylation techniques according to the type of DNA methylation measured (global or sequence-specific) and the principle of DNA methylation discrimination 48
Figure I.16. Chemical structures of major BFRs compounds 53
Figure I.17. Worldwide distribution of median PBDEs congeners indoor house dust concentrations 55
Figure I.18. Chemical structures of major PFASs compounds 58
Figure I.19. The occurrence of perfluoroalkyl acids in the global environment (including air, water, sediment and fish) 59
Figure I.20. Susceptibility windows of DNA-methylation due to environmental pollutants 62

Figure II.1. Number of new cancer cases attributable to lifestyle and environmental factors among adults aged 30 and over in France, 2015 76
Figure II.2. Mutation aetiology in lung adenocarcinoma 78
Figure II.3. Somatic mutation and stem cell division theories of cancer 79

Figure III.1. Barplot with the number of mutations in each sample in four TCGA cohorts. Each bar represents a sample, with the number of mutations shown in the y-axis. 95
Figure III.2. Simulations of 563 lung adenocarcinoma catalogues according to different models 97
Figure III.4. Reconstruction errors and their variability due to stochastic steps in the algorithms with and without pre-treatment to moderate the effect of hypermutated samples. 103
Figure III.5. Simulation study: specificity of extraction methods and mapping on COSMIC signatures as the number of analyzed catalogues and the cosine cut-off *h* vary. 105
Figure III.6. Simulation study: sensitivity of extraction methods and mapping on COSMIC signatures as the number of analyzed catalogues and the cosine cut-off *h* vary. 106
Figure III.7. Simulation study: specificity of extraction methods and mapping on COSMIC signatures as the average number of mutations and the cosine cut-off *h* vary 108
Figure III.8. Simulation study: sensitivity of extraction methods and mapping on COSMIC signatures as the average number of mutations and the cosine cut-off *h* vary 109

Figure III.9. Running times of *de novo* tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes. 110
Figure III.10. Simulation study: bias of the estimates of each signature contribution for several refitting methods. 111
Figure III.11. Running times of refitting tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes. 112

Figure IV.1. Calendar of self-administrated questionnaires in E3N..... 121
Figure IV.2. Organization of chips within plate..... 128
Figure IV.3. Correlation between the different BFRs congeners for blood concentrations 137
Figure IV.4. Quantile-quantile plot for the association between circulating levels of BFRs and DNA methylation at 805 837 CpGs sites (N=168) 138
Figure IV.5. Quantile-quantile plot for association between estimated dietary exposure to BFRs and DNA methylation at 805.837 CpGs sites (N=162)..... 139
Figure IV.6. Quantile-quantile plot for association between circulating levels of PFASs and dietary exposure to PFASand DNA methylation at 805.837 CpGs sites 150

LIST OF TABLES

Table I.1. Total contents in version 86 of the COSMIC database (August 2018). 40
Table I.2. Physicochemical properties of PBBs, PBDEs, and HBCDs 54
Table I.3. Physicochemical properties of PFOA and PFOS 58

Table II. Comparison between mutation rates, cumulative stem cell lifetime divisions, hazard ratios (HR) for cancer in smokers and mortality rates in smokers and never smokers, for the cancer sites for which information was available in all sources 81

Table III. Available tools for the detection of mutational signatures. 89

Table IV.1. Examples of random effects mixed-effects model formulas used in the lme4 R package. 129
Table IV.2. Baseline characteristics of the study population 135
Table IV.3. Distribution of BFRs concentrations in plasma (ng/g of lipids) and estimated dietary exposure to BFRs (ng/kg BW/day) in our study population (N=168 and N=162 respectively) 136
Table IV.4. Correlations between dietary exposure estimates and circulating levels of PBDEs congeners (N=162) 137
Table IV.5. Linear mixed effect models for circulating levels or dietary exposure to BFRs and genome-wide methylation M-value of 805 837 CpGs 141
Table IV.6. Linear mixed effect models for circulating levels of BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions 142
Table IV.7. Linear mixed effect models for dietary exposure to BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions 142
Table IV.8. Linear mixed effect models for circulating levels of BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 143
Table IV.9. Linear mixed effect models for dietary exposure to BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 143
Table IV.10. Gene set enrichment analysis results for genes that are positively or negatively correlated to BFRs exposure 145
Table IV.11. Baseline characteristics of the study population 148
Table IV.12. Distribution of PFASs concentrations in serum (ng/mL) and estimated dietary exposure to PFASs (ng/kg BW/day) in our study population (N=168 and N=162 respectively) 148
Table IV.13. Correlation between the different PFASs congeners for blood concentrations and estimated dietary exposure separately 149
Table IV.14. Correlation between dietary exposure estimates and circulating levels of PFASs congeners (N = 162) 149
Table IV.15. Linear model for circulating levels or dietary exposure to PFASs and genome-wide methylation of 805.837 CpGs 151

Table IV.16. Linear mixed effect models for circulating levels of PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 152
Table IV.17. Linear mixed effect models for dietary exposure to PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 152
Table IV.18. Linear mixed effect models for circulating levels of PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 152
Table IV.19. Linear mixed effect models for dietary exposure to PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions. 153
Table IV.20. Gene set enrichment analysis results for genes that are positively or negatively correlated to PFASs exposure (FDR < 0.3) 154

LIST OF APPENDICES

Appendix 1. Résumé en français 169
Appendix 2. MutationalCone implementation 183
Appendix 3. Overview of gene sets in MSigDB 184
Appendix 4. Description of hallmarks associated with BFRs or PFASs exposure 186
Appendix 5. Top 20 CpGs associated with dietary exposure to HBCDs congeners 187
Appendix 6. Top 20 CpGs associated with dietary exposure to PBDEs congeners 189
Appendix 7. Top 20 CpGs associated with circulating levels of PBDEs congeners 194
Appendix 8. Top 20 CpGs associated with circulating levels of PBB-153 197
Appendix 9. Top 20 CpGs associated with dietary exposure to PFASs congeners 198
Appendix 10. Top 20 CpGs associated with circulating levels of PFASs congeners 199

LIST OF ABBREVIATIONS

AA	Aristolochic Acid
AIMS	analysis of DNA methylation by amplification of intermethylated sites
AFB1	Aflatoxin B1
AHR	aryl hydrocarbon receptor
AHRR	aryl hydrocarbon receptor repressor
ANSES	agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail
AuNPs	Au nanoparticles
BFRs	Brominated Flame Retardants
BMI	Body Mass Index
BMIQ	beta-mixture quantile
Bps	base pairs
BS	bisulfite
BS-Seq	bisulfite sequencing
BSAS	bisulfite amplicon sequencing
CGIs	CpGs Islands
COBRA	combined bisulfite restriction analysis
COSMIC	Catalogue Of Somatic Mutations In Cancer
CNV	Copy Number Variation
CYP1B1	Cytochrome P450 Family 1 Subfamily B Member 1
CSC	cigarette smoke condensate
DNA	deoxyribose nucleic acid
E3N	Étude épidémiologique auprès de femmes de la MGEN
EDCs	Endocrine-Disrupting Chemicals
EFSA	European Food Safety Authority
ELISA	enzyme-linked immunosorbent assay
ENTPD2	Ectonuclease triphosphate diphosphohydrolase 2
EPA	Environnemental Protection Agency
FDR	false discovery rate
FRs	Flame Retardants
GD	gestational day
GSEA	Gene Set Enrichment Analysis
HBCD	Hexabromocyclododecane
HBM	Human Biomonitoring
HCC	hepatocellular carcinoma
HPF	hour post-fertilization
HPCE	high-performance capillary electrophoresis
IHC	immunohistochemistry
IL2	Interleukin 2
INCa	Institut National du Cancer
IPCS	International Programme on Chemical Safety
iPSC	induced pluripotent stem cell
kb	base pair
LC-MS	liquid chromatography coupled with mass spectrometry
LME	linear mixed effect
Log K_{OW}	octanol-water partition coefficient
LUMA	luminometric methylation assay

LSCD	Lifetime stem-cell divisions
MAF	Mutation annotation format
MBD-Seq	methyl-CpG binding domain sequencing
MCA	methylated CpG island amplification
MD	Mediterranean Diet
MeDIP	methylated DNA immunoprecipitation
MeDIP-Seq	methylated DNA immunoprecipitation sequencing
MFVF	Mutation Feature Vector Format
MM	Molecular Mass
MPF	Mutation Position Format
MOE	Margin Of Exposure
MS-AFLP	methylation-sensitive amplification length polymorphism
MSP	methyl-sensitive PCR
MYC	proto-oncogene
NF-κB	nuclear factor-kappa B
NGS	next-generation sequencing
NSUMA	next-generation sequencing of unmethylated Alu
NTL	non-tumour lung tissue
oxBS.Seq	oxidative bisulfite sequencing
PARK7	Parkinsonism associated deglycase
PBBs	polybrominated biphenyls
PBDEs	Polybrominated diphenyl ethers
PFOA	Perfluorooctanoic acid
PFOS	Perfluorooctanesulfonic acid
PFASs	Per-Fluorinated Alkylated Substances
PND	postnatal day
POPs	persistent organic pollutants
PCAWG	PanCancer Analysis of Whole Genomes
QUAlu	quantification of unmethylated Alu
RE	restriction enzyme
RRBS-Seq	reduced representation bisulfite sequencing
RNA	ribonucleic acid
RP-HPLC	reversed-phase high-performance liquid chromatography
RRBS	reduced representation bisulfite sequencing
SAM	S-Adenosyl-l-Methionine
SBS	Single Base Substitution
SCDTC	stem cell division theory of cancer
SMT	Somatic Mutation Theory
SNVs	Single Nucleotide Variants
STAT5	Signal transducer and activator of transcription 5
TAB-Seq	TET-associated bisulfite sequencing
TCGA	The Cancer Genome Atlas
TDS2	Second French Total Diet Study
TFs	Transcription Factors
TNF	Tumour Necrosis Factor
TSS1500	within 1500 bps of a transcription start site
TSS200	within 200 bps of a transcription start site
TSG	Tumour Suppressor Gene
TSL	total serum lipids
TSS	transcription start site
VCF	Variant Call Format
WGBS	whole genome bisulfite sequencing
WHO	World health organization

CHAPTER I:

GENERAL INTRODUCTION

This chapter serves as an introduction to most of the concepts discussed in my dissertation and will be divided into four sections, with the first three presenting background knowledge and recent advances about genomic and epigenomic signatures, and the last outlining the specific objectives and results of my thesis. Firstly, this introductive chapter will focus on genomics signatures, and in particular cancer mutational signatures, with a brief summary of concepts behind their definitions, mathematical modeling and identification. Next, we will discuss the best-studied epigenetic signatures, DNA methylation, focusing on methodological aspects and the influence lifestyle has on it. Finally, the third section will summarize current knowledge about brominated flame retardants and Per- and polyfluorinated alkylated substances, two classes of endocrine disrupting chemicals, and provide information about their impact on human health, as well as current developments in their molecular epidemiology.

This chapter does not review any of the articles that have been published or submitted as part of this thesis as these will be presented in the following chapters.

1. Genomic signatures 31
1.1 Behind the concept of "mutational signatures" 31
1.2 Mathematical modeling of a mutational process 35
1.3 COSMIC: Catalogue Of Somatic Mutations In Cancer 40
1.4 Experimental validation of mutational signatures 43
2. Epigenomic signatures 44
2.1 Introduction to epigenetics 44
2.2 Profiling DNA methylation 48
2.3 How does lifestyle influence DNA methylation 51
3. Endocrine disruptors 52
3.1 Introduction to Persistent Organic Pollutants 52
3.2 Persistent Organic Pollutants and DNA methylation 62
4. Summary and Objectives 64

1. GENOMIC SIGNATURES

1.1 BEHIND THE CONCEPT OF "MUTATIONAL SIGNATURES"

1.1.1 HALLMARKS OF CANCER

Living organisms are continuously exposed to a myriad of DNA damaging agents that can impact health and modulate disease-states[1] such as cancer which induce modifications in human genome resulting in an abnormal cell growth. In France, 382,000 new cases and 157,400 deaths have been observed in 2018[2].

Cancer encompasses more than 100 distinct diseases with diverse risk factors and epidemiology which originate from most of the cell types and organs of the human body and which are characterized by relatively unrestrained proliferation of cells that can invade beyond normal tissue boundaries and metastasize to distant organs[3]. This complexity points to a set of questions and investigations mainly related to regulatory mechanisms carcinogenesis that further lead to the identification of ten alterations in cell physiology that collectively dictate malignant growth and are shared by most and perhaps all types of human tumours[4].

Also known as "hallmarks of cancer", each of these physiologic changes represents novel capabilities acquired during tumour development and in particular the successful breaching of anticancer defense mechanisms hardwired into cells and tissues. These subsequent changes may explain why cancer is relatively rare during an average human lifetime. Six years later after the introduction of the original hallmarks, a revisited version consisting in seven categories was further proposed by Fouad and Aanei[5]. These hallmarks were defined as acquired evolutionary, advantageous characteristics that complementarily promote transformation of phenotypically normal cells into malignant ones and that promote progression of malignant cells while sacrificing/exploiting host tissue (Figure I.1).

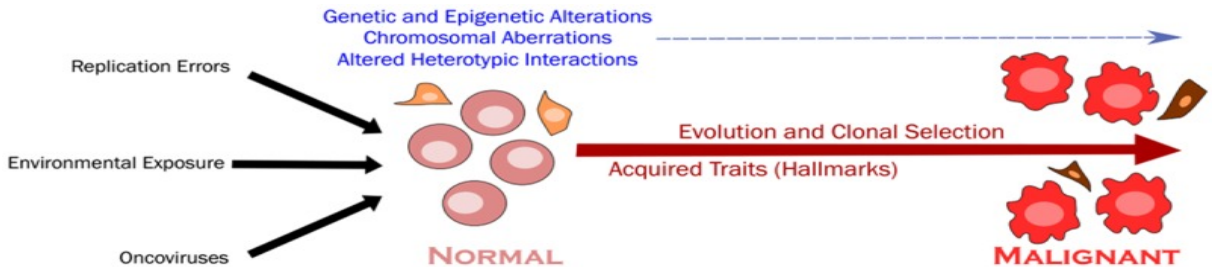


Figure I.1. **The transformation process of normal cells to malignant cells.**
Adopted from Fouad and Anei[5]

1.1.2 SOMATIC MUTATIONS AND RELATED THERORIES

Somatic mutations are defined as changes in the DNA sequence that are not passed on to the offspring through the germline[3]. Most current approaches in cancer research are based on Somatic Mutation Theory (SMT) that views somatic mutations as an epiphenomenon or a post-carcinogenesis event[5,6]. Briefly, cellular defects (mainly through to DNA damage) induce uncontrolled cell divisions that lead to the development of carcinogenesis suggesting that cancer is due to the accumulation of somatic mutations[7] (Figure I.2).

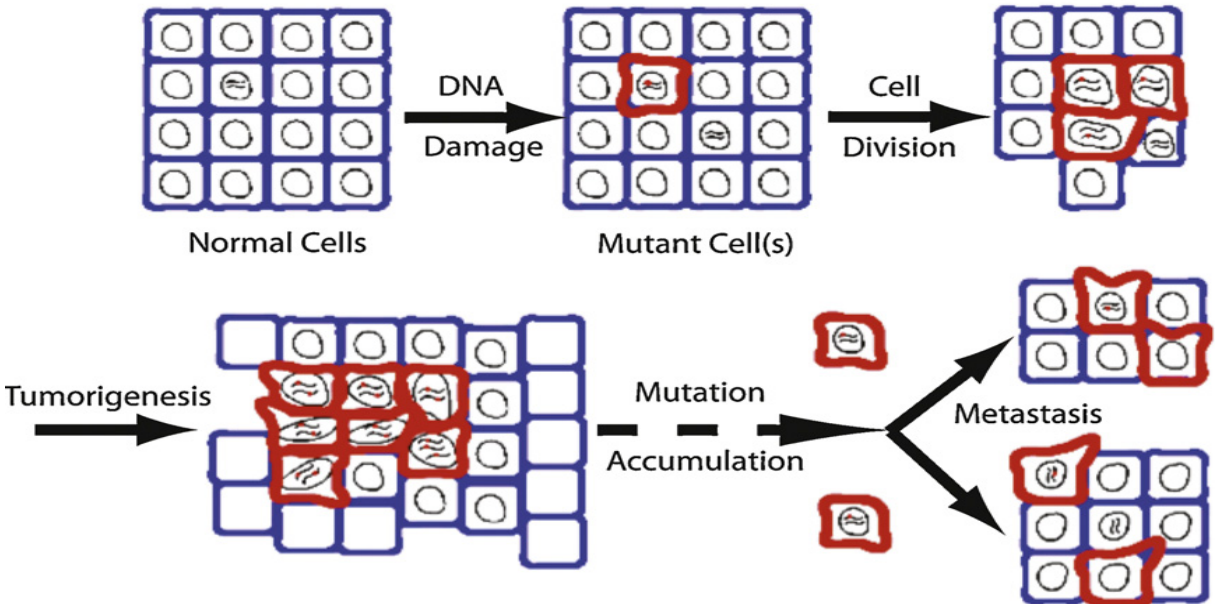


Figure I.2. **Somatic mutations leading to carcinogenesis**
Adopted from Kennedy and colleagues[7]

Historically, the SMT was first postulated in 1914 suggesting that a combination of chromosomal defects should result in cancer, followed by a proposal that mutations could cause cancer.

Two decades later, the understanding of the molecular structure of DNA lead to the 1-hit (mutation), 2-hit and hyper-mutation theories First, it was postulated that a person who inherits a mutant allele (1-hit) must experience a second somatic mutation (2-hit) to initiate carcinogenesis before further studies shown that for most cancer, more mutations are required (1953-2014). In 2007, they were categorized in two groups termed as “drivers”, those that confer a large selective advantage for tumour development and progression, and “passengers”, those that confer weaker selective advantage or are truly neutral in that they do not affect cancer cells’ survival.

Together, they both constitute a record of all cumulative DNA damage and repair activities occurred during the cellular lineage of the cancer cell[8]. A recent elaboration on the SMT was proposed in 2015 by Vogelstein and Tomasetti[9] who suggested that cancer development is an event that can be attributed to “bad luck” through accumulation of “enough” mutations that cause cancer.

This controversial claim will be discussed in chapter II and a summary of 100 years of research on the SMT can be found below[10] (Figure I.3).

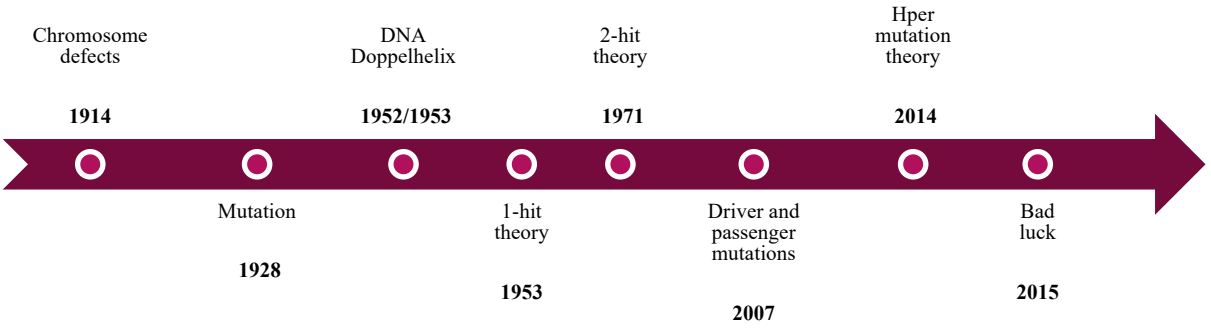


Figure I.3. **100 years of somatic mutations theory**
Modified from Brücher and Jamall[10]

1.1.3 BASE SUBSTITUTIONS AND GENOMIC ALTERATIONS

Cancer is a complex disease that involves mutant cells originating from a DNA modification in a single normal cell. Such modification is then propagated through cell divisions and accumulates with further DNA modifications finally leading to abnormal, cancerous cells[3]. Such somatic mutations include Single Nucleotide Variants (SNVs), insertions or deletions, Copy Number Variation (CNV) and chromosomal aberrations and are not to be confounded with those inherited and transmitted from parents (germline mutations). It is important to note that SNVs are different from SNPs (Single Nucleotide Polymorphisms). SNPs are single nucleotides substitutions expected to be present in a certain fraction of a given population and at the same position in both normal or cancer cells, while SNVs are only present in tumour cells and are likely shared in individuals with the same cancer.

As previously mentioned, somatic mutations can be endogenous, thus resulting from genome instability or deficiency in a DNA repair mechanism, or exogenous, that is due to environmental exposure such as tobacco smoking or UV light. For instance, UV light is known to induce DNA damage through C>T substitutions and could lead to a genotoxic stress that induces genome instability, while tobacco smoking induces T>A mutations.

With the development and the improvement of sequencing technologies collectively referred to as *High-Throughput Sequencing (HTS)* and the availability of cancer exome and genome data from most human cancers, much has been learnt about somatic mutations.

Among all of them, a particular focus has been placed on Single Base Substitutions (SBS) that have been classified in six types according to the mutated pyrimidine base (C or T) in a strand-symmetric model of mutation. Such 6 substitutions (C>A, C>G, C>T, T>A, T>C and T>G) may be further

classified in different types when considering the sequence pattern in which they are located (sequence context). For practical reasons, the sequence context is typically defined using the 5' and 3' bases proximal to the mutated base, that results in substitutions being classified in 96 types ($6 * 4 * 4$) (Figure I.4).

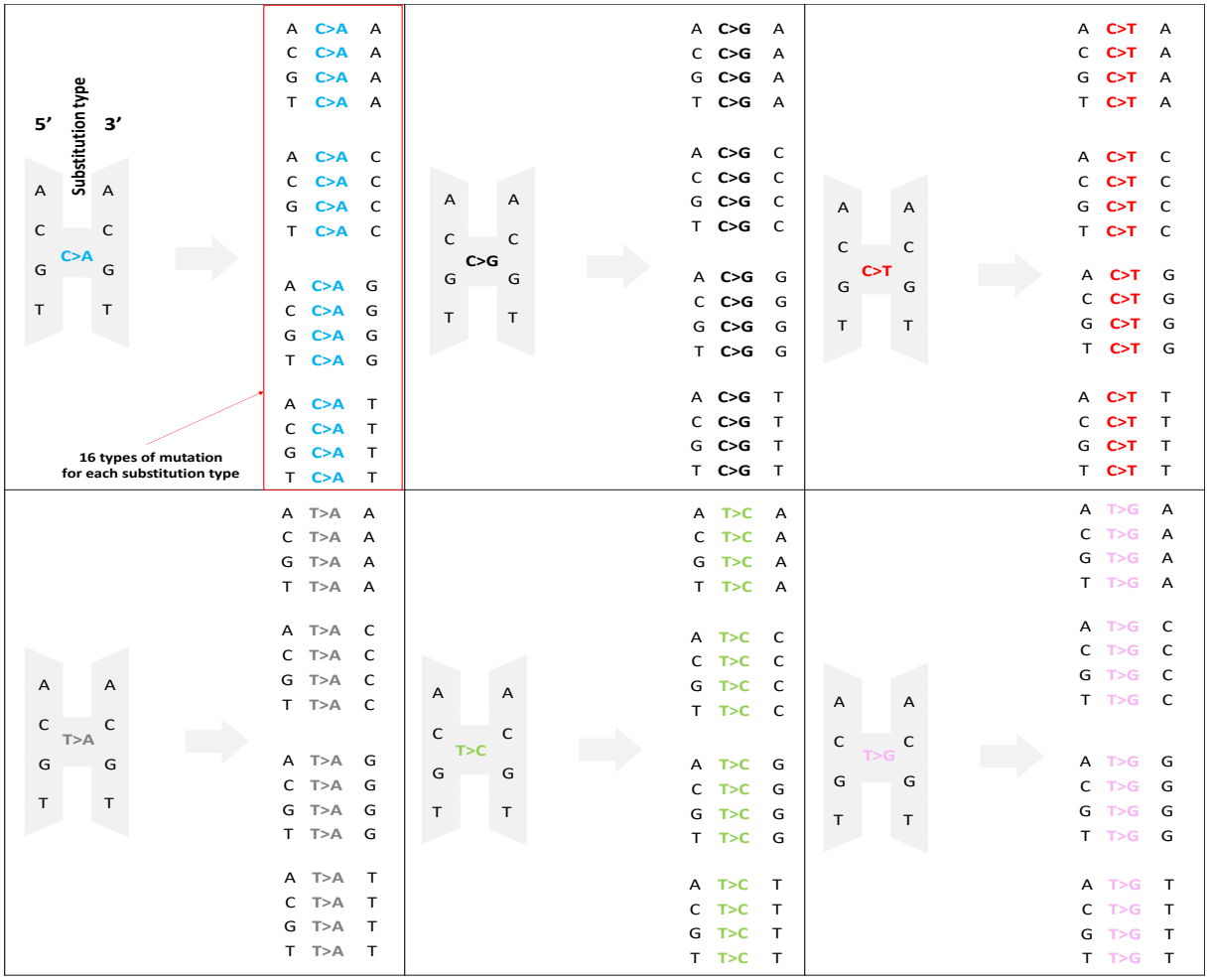


Figure I.4. **The 96 mutations types in a trinucleotide context**
Considerations of the 6 types of base substitutions_ a DNA base is replaced by another (C>A, C>G, C>T, T>A, T>C and T>G) and the associated sequence context.

It has been hypothesized that mutational processes leave specific patterns of somatic mutations, so-called mutational signatures. To identify such patterns from the substitutions measured from cancer samples, computational models, such as matrix decomposition algorithms or probabilistic models, have been developed. The first of such methods was published in 2013 by Alexandrov and colleagues[11], and, as for most of all the other models that followed, is based on the idea that a mutational signature can be seen as a probability distribution of the 96 types of mutations or more according to the length of the sequence context. Mutational signatures contribute to the total mutational burden of a cancer genome, commonly referred to as mutational "catalogue" or "spectrum" in the recent computational biology literature.

1.2 MATHEMATICAL MODELING OF A MUTATIONAL PROCESS

1.2.1 DEFINITION OF MUTATIONAL CATALOGUES, SPECTRA AND SIGNATURES

The mutational catalogue representing the total mutational burden of a genome (or exome) g is defined as a vector $(m_g^1, \dots, m_g^K)^T$, where each m_g^k is the number of mutations of type k found in the genome and K, the number of possible mutation types, is equal to 96. The superscript T denotes the transpose of a matrix so that vectors are thought as column vectors. In this setting, information about mutation locations in the sequence is lost and the catalogue is built by comparing the sequence to a reference sequence in order to detect mutations and then by simply counting the occurrences of each type. The reference sequence can either be a standard reference (e.g. the assembly GRCh38 of 2013 also known as hg38 or the previous one GRCh37 with reference to hg19) or a sequence from a "normal" tissue from the same individual (e.g. DNA from blood or from normal tissue surrounding tumours when available).

For the purposes of the present thesis, the generic term "samples" will be used for both genomes and exomes as the concepts and models used may be applied to both.

The basic idea underlying all computational models proposed is that the mutational catalogue of a sample results from the combination of all the mutational processes operative during lifetime, and therefore it can be seen as the weighted superposition of simpler mutational *signatures,* each uniquely corresponding to a specific process. The weight is larger if the process has a larger role in the final catalogue of mutations: for example, mutagens that last longer, are more intense, generate poorly repaired DNA lesions, mutate more genes, or also act as selection pressures favoring mutant cells.

Formally, the signature of a mutational process n is a vector $p_n = (p_n^1, \dots, p_n^K)^T$, where each p_n^k represents the probability that the mutational process will induce a mutation of type k. In other words, p_n^k is the expected relative frequency of type k mutations in genomes exposed to n.

Note that $\sum_{k=1}^{K} p_n^k = 1$ and $0 \leq p_n^k \leq 1$ for all k.

The intensity of the exposure to a mutational process n in a sample g is measured by the number of mutations e_g^n in g that are due to n. For this reason, e_g^n is referred to as the "exposure" of g to n. It is important to notice that the term "exposure" does not refer here to the exposure to a mutagen *per se*, because it also includes the likelihood that an unrepaired DNA lesion will cause a mutation. The expected number of mutations of type k due to the process n in sample g is therefore $p_n^k e_g^n$. If sample g has been exposed to N mutational processes, then the total number of mutations of type k is :

$$m_g^k = \sum_{n=1}^{N} p_n^k \, e_g^n + \epsilon_g^k, \qquad (1)$$

where ϵ_g^k is an error term reflecting sampling variability and non-systematic errors in sequencing or subsequent analyses.

Matrix notation is effectively used when dealing with several samples and signatures. In this situation, the collection of G samples is represented by the $K \times G$ matrix, with catalogues in columns:

$$M = \begin{pmatrix} m_1^1 & m_2^1 & \dots & m_G^1 \\ \vdots & \vdots & & \vdots \\ m_1^K & m_2^K & \dots & m_G^K \end{pmatrix}, \text{ Figure I5.A)}$$

the N signatures are represented by the $K \times N$ matrix

$$P = \begin{pmatrix} p_1^1 & p_2^1 & \dots & p_N^1 \\ \vdots & \vdots & & \vdots \\ p_1^K & p_2^K & \dots & p_N^K \end{pmatrix}, \text{ Figure I.5.B)}$$

and the exposures by the $N \times G$ matrix

$$E = \begin{pmatrix} e_1^1 & e_2^1 & \dots & e_G^1 \\ \vdots & \vdots & & \vdots \\ e_1^N & e_2^N & \dots & e_G^N \end{pmatrix}. \text{ Figure I.5.C)}$$

Equation (1) then becomes : $M \approx P \times E$ *where we omitted the error term.*

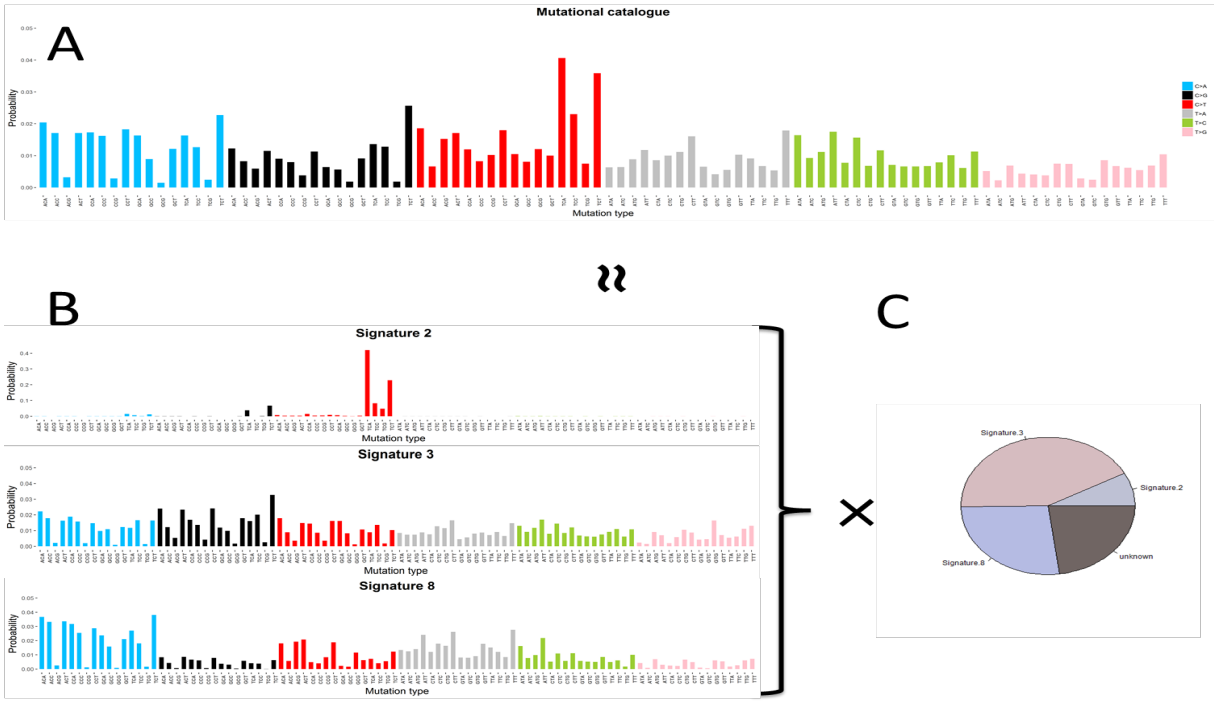


Figure I.5. **Mutational catalogue and the individual signatures contribution to it**
A) Mutational catalogue of a breast cancer genome PD4107a[12]. B) The catalogue is the result of the linear combination of COSMIC signatures 2, 3 and 8 with some additional noise. C) Relative burden of each signature.

1.2.2 DECIPHERING THE SIGNATURES OF MUTATIONAL PROCESSES: *DE NOVO VS. REFITTING*

De novo signature extraction methods aim at estimating P and E given M. Non-negative matrix factorization (NMF) is an appealing solution to this unsupervised learning problem, because, by definition, all involved matrices are non-negative. NMF was popularized in 1999 by Lee and Seung and has become a widely used tool for the analysis of high dimensional data, mainly image processing or recognition and text mining.

In the context of mutational signatures, NMF identifies two matrices P and E that minimize the distance between M and $P \times E$. In particular, NMF finds an approximated solution to the non-convex optimization problem:

$$argmin_{P \geq 0,\, E \geq 0} ||M - P \times E||_F^2, \quad (2)$$

where the Frobenius matrix norm of the error term is considered.

We recall that the Frobenius norm of a matrix is simply the square root of the sum of the squares of all the matrix elements.

NMF requires the number of signatures N, an unknown parameter, to be predefined or estimated. An approach for selecting this parameter consists in obtaining a factorization of M for several of its values and then choosing the best N with respect to some performance measure such as the reconstruction error or the overall reproducibility. NMF is at the core of the Wellcome Trust Sanger Institute (WTSI) Mutational Signature Framework, the first published method for signature extraction[11]. An alternative to numerical approaches based on NMF is given by statistical modelling and algorithms. With these latter approaches, the number of mutations of a given type can be modelled by a Poisson distribution

$$m_g^k \sim \mathcal{P}\left(\sum_{n=1}^{N} p_n^k \, e_g^n\right)$$

where mutational processes are assumed to be mutually independent.

This latter independence hypothesis simplifies the mathematics but does not necessarily hold in practice, where mutation processes are likely to interfere with each other (e.g. distinct defective DNA repair processes). In order to estimate E and P, it has been proposed to consider E as latent data and P as a matrix of unknown parameters and to apply an expectation-maximization algorithm[13] or use Bayesian approaches[14]. One important advantage of statistical approaches is the availability of model selection techniques for the choice of N.

The refitting approaches consider that the signatures P are known and the goal is to estimate E given M and P. Refitting can be done for individual mutational catalogues (i.e. individual samples) and, from a linear algebra perspective, can be seen as the problem of projecting a catalogue living in the K-dimensional vector space (the space spanned by all mutation types) onto its subset of all linear

combinations of the given mutational signatures having non-negative coefficients (the cone spanned by the given signatures).

A current practice consists in first performing a de novo extraction of signatures followed by a comparison of the newly identified signatures with the reference signatures (e.g. the COSMIC signatures introduced in the next section) by means of a similarity score, typically cosine similarity ranging from 0 (completely different) to 1 (identical)[10,11]. A "novel" signature is considered to reflect a specific reference signature if the similarity is larger than a fixed cut-off. If similarity is observed with more than one reference signature, the one with the largest value of similarity is chosen (Figure I.6).

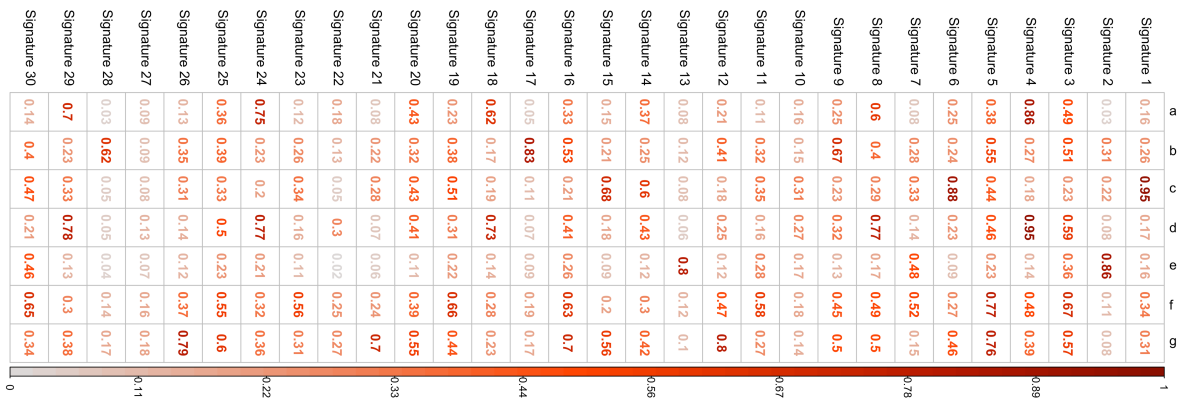


Figure I.6. **Comparison of newly identified signatures with COSMIC signatures**

Signatures *a*-*g* were identified in a de novo extraction using the maftools[16] R package from the The Cancer Genome Atlas lung adenocarcinoma cohort which include 563 cancer genomes at the date of selection. The novel signatures were then compared to the 30 signatures validated in the COSMIC database in terms of cosine similarity. Each signature is then assigned to the most similar COSMIC signature provided that their cosine similarity is above a fixed threshold. For instance, signature f is matched to signature 5 at a cut-off of 0.75 but is considered as a completely new signature if the cut-off is at 0.80. Also note that a unique assignment can be controversial: for instance, signature g is similar both to signatures 12 and 26 (Figure I.7).

This assignment step crucially depends on the choice of the cut-off *h* that has been so far inconsistent in the literature with some studies using a value of 0.75[17] whereas others 0.80[18,19]. Another difficulty is that different signatures might have very close cosine similarity, as it happens also between COSMIC signatures, so that a unique assignment is not always possible. This shows that mutational signatures are a useful mathematical construct that, however, might have biological ambiguous meaning.

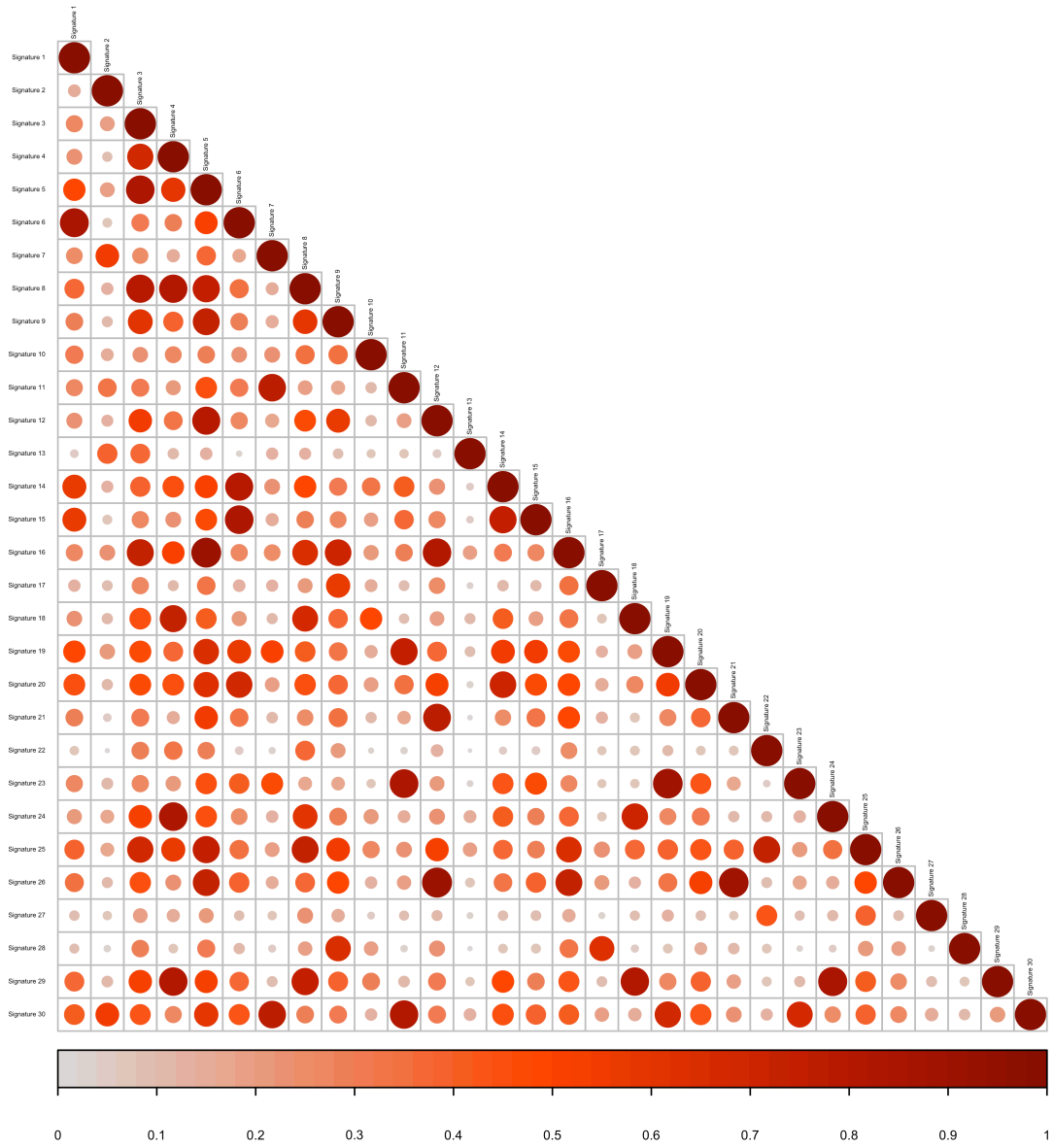


Figure I.7. **Cosine similarity plot of COSMIC signatures**

1.3 COSMIC: CATALOGUE OF SOMATIC MUTATIONS IN CANCER

The Catalogue Of Somatic Mutations In Cancer (COSMIC) available at http://cancer.sanger.ac.uk/cosmic/signatures, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. Built in 2004, the database and website have been developed to store somatic mutation data in a single location and display the data and other information related to human cancer.

In addition to coding mutations, COSMIC covers all the genetic mechanisms by which somatic mutations promote cancer (Figure I.8). In parallel, the Cancer Gene Census (CGC) describes a curated catalogue of genes driving every form of human cancer using the ten hallmarks as proposed by Hanahan and Weinberg[4].

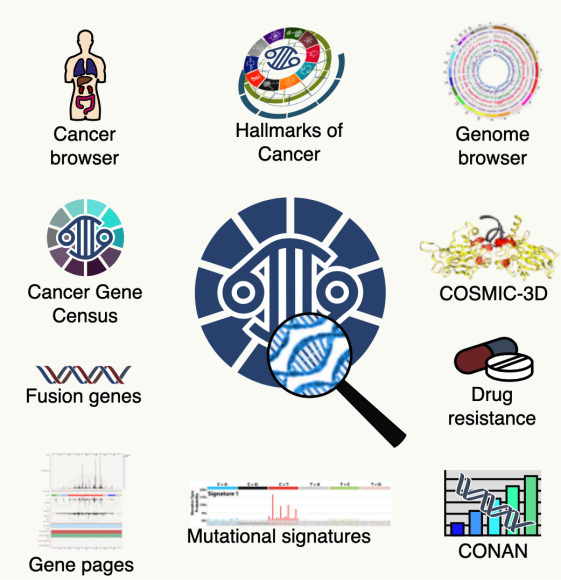


Figure I.8. **Overview of COSMIC tools**
Adopted from COSMIC

Data within COSMIC are updated constantly and released on a regular, three-monthly cycle, guaranteeing four releases per year[20]. As example, one of the last updates (Table I.1, August 2018) includes almost 6 million coding mutations across 1.4 million tumour samples.

Table I.1. Total contents in version 86 of the COSMIC database (August 2018).
Adopted from Tate and colleagues[20]

1 391 372	Tumour samples
5 977 977	Coding Mutations
26 251	Manually Curated Publications
19 368	Gene Fusions
35 480	Whole Genomes/Exomes across **457** studies/papers
1 179 545	Copy Number Variants
9 147 833	Gene Expression Variants
7 879 142	Differentially Methylated CpGs
19 721 019	Non-coding Variants

The application of the mutational signature's framework to tens of thousands of genomes and exomes from 40 different cancers types from large data repositories such as TCGA (The Cancer Genome Atlas), has led to the identification of 30 mutational signatures (Figure I.9) characterized by a unique probability profile across the 96 mutation types. These validated mutational signatures are listed in a repertory on the COSMIC website and have been widely used as references (Mutational signatures v2).

Figure I.9. **Patterns of mutational signatures (v2 – March 2015): 30 SBS**
Adopted from COSMIC

More recently, Alexandrov et al. have introduced an updated set of signatures identified from an even larger collection of both exome and whole-genome sequences (including the sequences from the PanCancer Analysis of Whole Genomes also known as PCAWG project) using two different methods (a new version of the original framework and a Bayesian alternative[21]). The new repertory includes 49 mutational signatures (Mutational signatures v3, Figure I.10) based on SBS as in the previous version, and also mutational signatures built in the context of other types of mutations such as Double Base Substitutions or DBS (11 signatures), clustered based substitutions (4 signatures) and small insertions and deletions (17 signatures).

Figure I.10. **Patterns of mutational signatures (v3 – May 2019) : 49 SBS**
Adopted from COSMIC

1.4 EXPERIMENTAL VALIDATION OF MUTATIONAL SIGNATURES

Since the publication of the first work about mutational signatures in 2013[11], multiple algorithms have been developed, leading to similar but not identical results, a source of concern for researchers interested in this type of analysis. Conceptually, this is not surprising: mutational signatures are naturally defined in terms of non-negative matrix factorization, a well-known ill-posed problem (a unique solution does not exist). Although this limitation has cast doubts on the biological validity of mutational signatures, this has been somehow validated using experimental and computational approaches by Zou and colleagues[22]. Sufficiently detailed tumour catalogues and mutagen spectra might yield patterns that are unique to a tumour type or mutagen, and therefore become "true" signatures that allow backward inference from the tumour to the mutagen. Mutational signatures data in combination with epidemiological information may provide useful insights to identify the causes of cancer[23,24]. The utility of the current models of substitution mutational signatures is also shown in a recent experimental work based on a human induced pluripotent stem cell (iPSC) line that provides evidence for the possibility to identify the agents responsible for some specific mutational signatures[25]. In such work, Kucab and colleagues compared iPSCs treated and untreated with 79 known or suspected environmental carcinogens and identified specific substitution mutational signatures for around half of such carcinogens. Some of such signatures were similar to those identified in human tumour DNA.

2. EPIGENOMIC SIGNATURES

2.1 INTRODUCTION TO EPIGENETICS

2.1.1 OVERVIEW

The word "epigenetics" literally means "in addition to changes in the genetic sequence" [26]. Epigenetics thus encompasses a wide range of mechanisms at the molecular level that can influence gene expression without involving changes to the underlying DNA sequence. As a matter of fact, even if every cell in a given individual contains the same DNA sequence, the molecular pattern leading to gene expression and protein synthesis is different. For instance, brain and lung cells are characterized by different physiological mechanisms and thus require different patterns of gene expression.

Reflecting how cells translate the information contained in the genetic sequence, are common to many organisms and is essential to their physiological functions. Aberrant modifications of epigenetic processes may have major adverse health and behavioral effects. Indeed, one of the most interesting fact of epigenetics is that its marks or states in cells change in response to outside influences. Studying epigenetic processes may therefore be helpful in addressing key questions such as: why are some foods good for our health while others are unhealthy particularly for groups of individuals? How does physical activity exert beneficial effects on several health outcomes? How do particular environmental exposures or psycho-social stress exert their detrimental effects on health?

Epigenetics is essentially additional information layered on top of the genetic sequence of the four nucleotides that makes up our DNA. Important modifications are the addition of molecules (methyl groups) or proteins (called histones) to the DNA sequence. Sometimes, epigenetic modifications are stable and passed on to future generations. Though DNA sequence is fairly permanent, and as previously mentioned, epigenetic modifications in other instances are dynamic and change in response to environmental stimuli. Thus, epigenetic is the study of mitotically heritable yet potentially reversible, molecular modifications to DNA and chromatin without alteration to the underlying DNA sequence[27].

There are multiple epigenetics mechanisms that may play a role in gene regulation machinery but the most studied and well-known remain histone modifications and DNA methylation. These are two process crucial to normal development and differentiation of distinct cell lineages in the adult organism, that if modified by exogeneous influences, and, as such, can contribute to or be the result of environmental alterations of phenotype or pathophenotype[28]. Other modifications include RNA

regulations, such as long non-coding RNAs that play an essential role in imprinting and X-chromosome inactivation or small non-coding RNAs known for their effects on transcriptional gene silencing.

Today, a wide variety of illnesses, behaviors, and other health indicators already have some level of evidence linking them with epigenetic mechanisms, including cancers of almost all types, cognitive dysfunction, and respiratory, cardiovascular, reproductive, autoimmune, and neurobehavioral illness[26]. Also, it is increasingly recognized that epigenetic marks (methylation cytosines residues on DNA, post-translational modification of histone tails and microRNA expression) provide a mechanistic link between environment, nutrition and disease.

2.1.2 DNA METHYLATION AND EPIGENETIC MECHANISMS

Molecular mechanisms of DNA methylation

From a molecular point of view, DNA methylation is a biochemical process that refers to the catalytic addition of a methyl (-CH_3) group to the fifth carbon position of a DNA base, usually a cytosine residue that is followed on the same strand by guanine, what is also known as CpG site (Figure I.11). In human genomes, CpGs dinucleotides are asymmetrically distributed and often concentrated in dense regions mostly unmethylated, called CpGs Islands (CGIs) that span the promoter of approximately one-half of all genes[29].

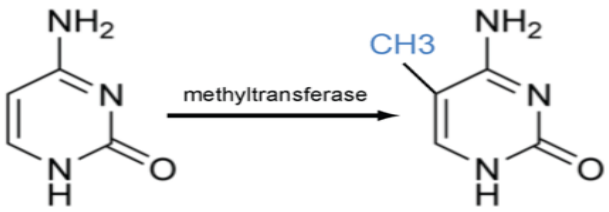


Figure I.11. **DNA methylation**
Credits to LabRoots

Approximately 80% of CpG dinucleotides outside of promoter regions are methylated under normal physiologic circumstances. Genome-wide decreases in methylation, or hypomethylation, are most functionally relevant when they occur in coding regions of genes, leading to alternative versions or levels of messenger RNA. In the other hand, the addition of methyl groups, or hypermethylation, can be highly specific to a particular gene with hypermethylation of CpG islands in the promoter region of a gene, known to result in transcriptional silencing of the gene, and subsequent loss of protein expression[30].

The enzymes that play a key role in methylation processes are called the DNA methyltransferases (DNMTs), with three of them DNMT1, DNMT3a and DNMT3b responsible of the establishment of

DNA methylation by catalyzing the transfer of a methyl group by the primary methyl donor named S-Adenosyl-l-Methionine (SAM) (Figure I.12).

DNMT1 is the most abundant methyltransferase in somatic cells and is responsible for the maintenance of DNA methylation during DNA synthesis for copying the original DNA methylation pattern to the newly formed strands. DNMT3a and DNMT3b are known to perform *de novo* methylation during embryonic development.

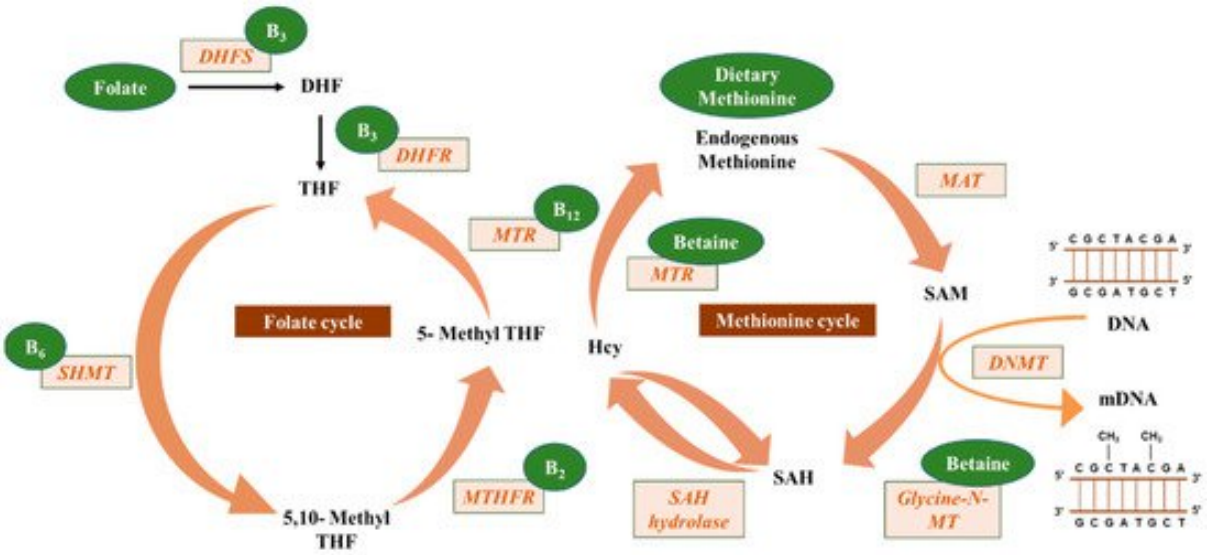


Figure I.12. **Micronutrient donors involved in one-carbon metabolism and subsequently in DNA methylation (one-carbon metabolism)**
Adopted from Mahmoud and Ali[31]

The role of DNA methylation

Over the last decades, several discoveries have been made about DNA methylation and how important it is for a number of cellular or developmental processes including embryonic development, X-chromosome inactivation, genomic imprinting, gene suppression, carcinogenesis and chromosome stability by silencing repetitive elements, and in maintaining tissue-specific and appropriate patterns of gene expression through cell division[32–34].

One major role of DNA methylation related to genome stability is structural and involves chromosomal and chromatin structure. Chromatin is a complex of DNA and proteins localized in the nucleus of eukaryotic cells that play major roles in various metabolic processes such transcription, replication or DNA repair. Chromatin can be divided into euchromatin and heterochromatin. As an example, alterations of heterochromatin through global hypomethylation is known to be a prerequisite for genome instability, which has been frequently reported to be associated with aging[35,36] (mainly due to telomeric chromosomal regions that represent regions of repetitive nucleotides at the end of chromosomes, known

to be a hallmark of senescence[37]) and certain pathology such as cardiovascular[38,39] or neurodegenerative[40,41] diseases and cancer[42].

Traditionally, cancer has been viewed as a disease driven by accumulation of mutations with this paradigm now expanded to incorporate disruption of epigenetic regulatory mechanisms[43]. As example, studies on molecular mechanisms underlying the role of DNA methylation in gene expression identified how epigenetic DNA modifications modulate the Transcription Factors (TFs) binding site to DNA for activation or repression of transcription (Figure I.13). It is now known that mutations on Tumour Suppressor Genes (TSG) or oncogenes (genes that can potentially lead to cancer) cause either loss or gain of function and abnormal expression. TSGs are genes usually silenced in cancerous cells due to hypermethylation in their promoter region and it is widely accepted that this phenomenon lead to tumourigenesis[44]. In a translational approach, hypermethylation of CpG promoter which is visible during early stages of some cancers such as colon cancer has the potential to serve as a biomarker of the disease[45].

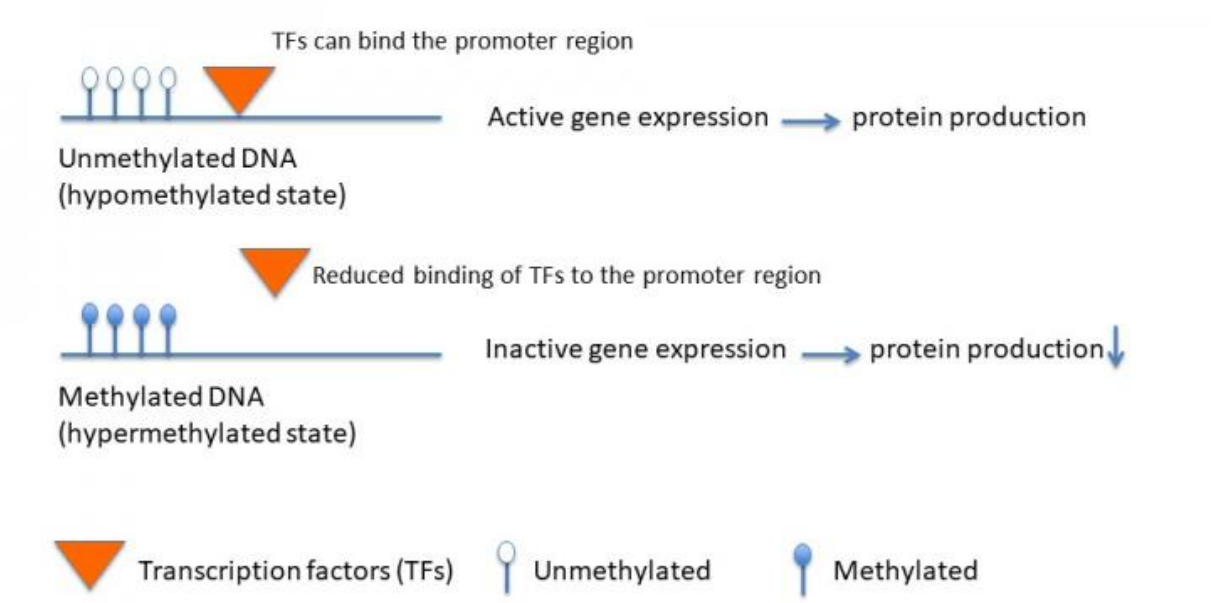


Figure I.13. **Effect of DNA methylation on gene expression**
Credits to Daniela Furrer, Laval University

2.2 PROFILING DNA METHYLATION

2.2.1 METHODOLOGICAL ASPECTS

Methods to analyze genome-wide DNA methylation patterns is still evolving and a wide range have been developed to generate quantitative and qualitative information on DNA methylation (Figure I.14).

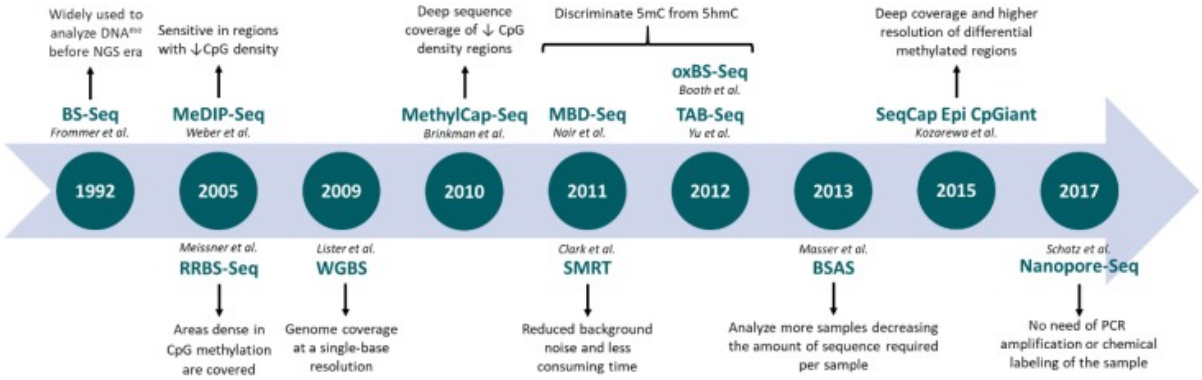


Figure I.14. **Evolution of next-generation sequencing-based techniques applied to DNA methylation profiling.**
Adopted from Barros-Silva and colleagues[46]

Generally, all of the methods include two procedures: the methylation-dependent pretreatment (including enzyme digestion, affinity enrichment or bisulfite conversion[47]) of the DNA and the following analytical step.

Then, the methods can be viewed according to the type of DNA methylation measured (global or sequence-specific) and the pre-treatment (Figure I.15).

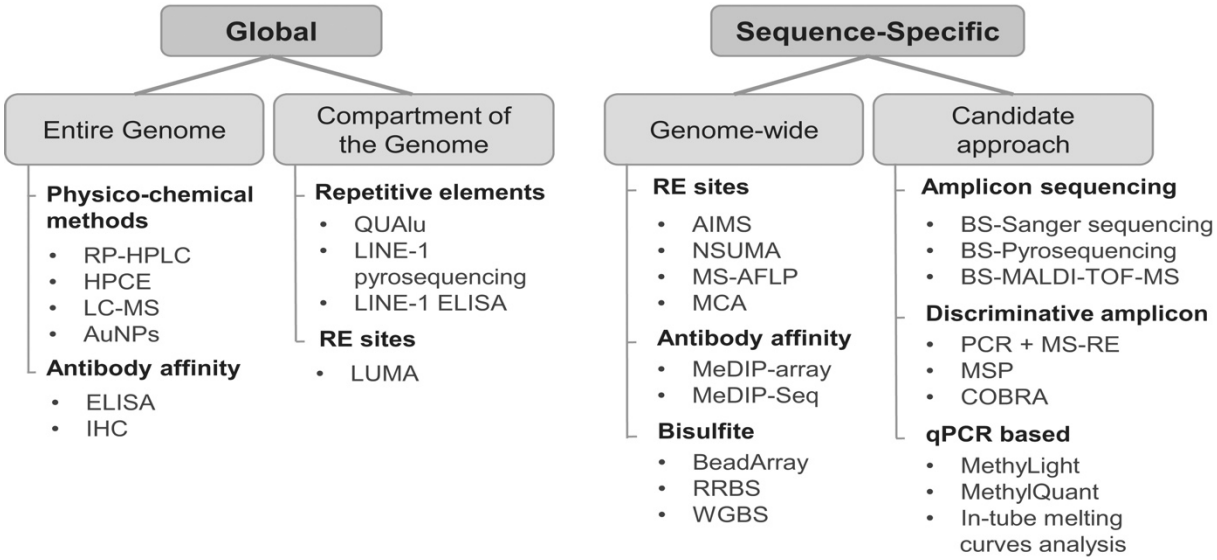


Figure I.15. **Main DNA methylation techniques according to the type of DNA methylation measured (global or sequence-specific) and the principle of DNA methylation discrimination**
Adopted from Zafon and colleagues[48]

Methods related to global methylation can be subdivided into those measuring the DNA methylation of the entire genome and those measuring the DNA methylation of a compartment of the genome used as surrogate reporter of the genome (e.g., repeat sequences such as LINE-1 and Alu elements, which comprise 20% and 10% of the human genome, respectively). Sequence-specific methods can also be subdivided into those that are genome-wide (mostly based on bead arrays or NGS) and those measuring specific regions of interest (mostly based on polymerase chain reaction)[48].

Recently, with the third-generation sequencing (Nanopore-Seq), sequencers allow for direct read of different modifications on DNA bases without DNA amplification or chemical labelling. Although these technologies are still in the development phase, they seem promising for future methylome profiling analysis.

The array-based methods and specifically the Illumina EPIC array used in the studies presented in the second part of the thesis, are methods based on bisulfite conversion of DNA and fall under the category "BeadArray".

2.2.2 BETA-VALUES AND M-VALUES IN MICROARRAY ANALYSIS

The microarray-based Infinium methylation assay by Illumina is one platform for low-cost high-throughput methylation profiling. Briefly, to estimate the methylation status, the Illumina Infinium assay utilizes a pair of probes (a methylated probe and an unmethylated probe) to measure the intensities of the methylated and unmethylated alleles at the interrogated CpG site. The methylation level is then estimated based on the measured intensities of this pair of probes.

To date, two methods have been proposed to measure the methylation level. The first one is called Beta-value, ranging from 0 to 1, which has been widely used to measure the percentage of methylation. The Beta-value is the ratio of the methylated probe intensity over the overall intensity (sum of methylated and unmethylated probe intensities) and is defined using the following formula:

$$Beta_i = \frac{\max\left(y_{i,methy}, 0\right)}{max\left(y_{i,unmethy}, 0\right) + \max\left(y_{i,methy}, 0\right) + \alpha}$$

where $y_{i,menty}$ *and* $y_{i,unmenty}$ *are the intensities measured by the* i^{th} *methylated and unmethylated probes, respectively.* α *is a constant offset and is generally equal to 100.*

The second method is the log2 ratio of the intensities of methylated probe versus unmethylated probe as shown in the following equation:

$$M_i = log_2\left(\frac{\max\left(y_{i,methy}, 0\right) + \alpha}{max\left(y_{i,unmethy}, 0\right) + \alpha}\right)$$

M-values are related to beta-value through the following logit transformation:

$$Beta_i = \frac{2^{M_i}}{2^{M_i} + 1}; M_i = log_2(\frac{Beta_i}{1 - Beta_i})$$

Beta-values have a more intuitive biological interpretation (it corresponds roughly to the percentage of a site that is methylated) but their distribution is not normal and is not homoscedastic (for high and low values of betas, the standard deviation is lower than for intermediate values). The distribution of M-values is closer to the normal and it is homoscedastic. Thus, M-values are therefore to be preferred for example in linear regression when methylation is the dependent variable.

2.3 HOW DOES LIFESTYLE INFLUENCE DNA METHYLATION

The property of environmental factors to induce epigenetics modifications highlight how and why monozygotic twins are not completely identical.

Exposure and lifestyle factors that modify the human epigenome are referred to as "epigenetic agents" and include behaviors, nutrition, chemicals and industrial pollutants that result in distinct gene expression profile. For example, nutrition is a key environmental exposure from gestation to death that impacts our health by influencing epigenetic phenomena. Recent epidemiological data suggest that the increased incidence of cancer observed in the developed world since the 1960s may partly be due to exposure to Endocrine-Disrupting Chemicals (EDCs), to which humans and wildlife are exposed daily from multiple sources[49]. The implication of other epigenetic agents such as tobacco, alcohol and obesity, in multifactorial diseases have been addressed through epidemiological studies that have shown association between gene-specific DNA methylation patterns and cancer incidence[31,50–52].

Smoking is a major risk factor for tobacco related cancers and many studies have been conducted in order to identify functional consequences of tobacco exposure and tobacco-related cancers metabolic alterations. Altered methylation levels in thousands of CpG sites have been found to be associated with smoking and smoking duration and intensity[53]. In case–control studies nested within prospective cohorts, some of these alterations have been found to be associated with lung-cancer risk even after adjustment for reported history of cigarette smoking[54].

With regards of the impact of diet on DNA methylation, and with consideration of one-carbon metabolism, it has been reported that diet containing high concentrations of choline and betaine is associated with reduced breast cancer mortality[55] and primary liver cancer[56]. Strong evidence shows that a dietary pattern inspired by Mediterranean Diet (MD) principles is associated with numerous health benefits, by increasing life expectancy with mainly protective effects on cardiovascular diseases and certain types of cancer[57]. The MD is not only a dietary pattern but also embodies social behavior and a way of life. Although different countries in the Mediterranean region have their own diets, they share the following pattern such as high consumption of extra virgin olive oil, legumes and nuts, unrefined cereals, fruits and vegetables, moderate consumption of dairy products, mainly cheese or yogurt, fish and wine and low consumption of meat and meat products. As DNA methylation is modulated by diet, a few studies investigated whether adherence to MD is associated with changes in DNA methylation from peripheral blood cells with results suggesting that MD is associated with changes in the epigenome[58].

However, "nutritional epigenetics" is a recent field of interest and the current knowledge about the precise effects of bioactive food components on epigenome and their potential association with the phenotype is limited.

3. ENDOCRINE DISRUPTORS

Endocrine Disrupting Chemicals (EDCs) are "exogenous substances or mixtures that alter the function(s) of the endocrine system, causing adverse health effects in an intact organism, its progeny, or (sub)populations"[59]. Such broad class of chemicals includes a variety of substances that are produced through components such as industrial solvents, food packaged, commercial household products (including stain- and water-repellent fabrics, polishes, waxes, paints, cleaning products), workplace (production facilities or industries such as chrome plating, electronics manufacturing or oil recovery) and that are released in the environment.

The effect of such substances on biological systems and their widespread presence in the environment, including in food, have led to growing concerns about the impact of EDC exposure on population health in industrialized countries. EDCs were indeed identified as "Substances of Very High Concern" by the Regulation (EC) No 1907/2006 of the European Parliament but the assessment of the health effects of specific EDCs is complex due to the vast number of such substances and their heterogeneity. In this research project we will focus on Brominated Flame Retardants (BFRs) and Per- and polyfluoroalkyl substances (PFASs), two classes of the broad group of EDCs called Persistent Organic Pollutants that have the characteristic of persisting in the environment for a long period of time and may therefore pose a hazard to human health.

3.1 INTRODUCTION TO PERSISTENT ORGANIC POLLUTANTS

Persistent Organic Pollutants (POPs) are EDCs of global concern due to their potential for long-range transport, persistence in the environment, ability to biomagnify and bioaccumulate in ecosystems that means they gradually accumulate in living organisms, as well as their action on the environment, on biological systems and in humans and other animals. Humans are widely exposed to these chemicals in a variety of ways but, due to their bioaccumulation, the most important route is through diet and, in particular, the consumption of foods of animal origin. POPs can also be found in the air and products used in our daily lives such as pesticides or solvents. Exposure to POPs can increase cancer risk, may lead to reproductive disorders, and some of these substances may increase the risk of birth defects through their genotoxic action.

Due to their bioaccumulation in the environment and the corresponding effect on human health, the international community has called for actions to reduce and eliminate production, use and releases of these substances through two international legally binding instruments:

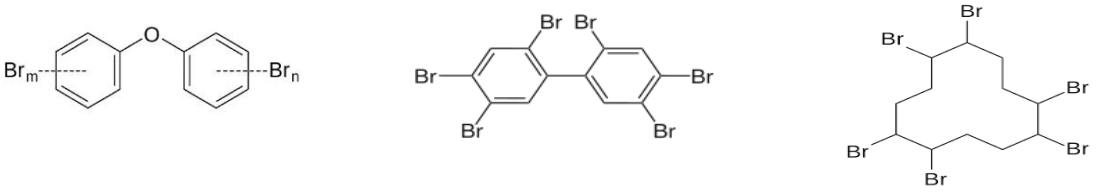
- ➔ The global Stockholm Convention on POPs, opened for signatures in May 2001 and entered into force on 17 May 2004;
- ➔ The Protocol to the regional UNECE Convention on Long-Range Transboundary Air Pollution (CLRTAP) on POPs, opened for signatures in June 1998 and entered into force on 23 October 2003.

BFRs and PFASs are two large families of environmental EDCs, for which the long-term health effects remain unclear and not well characterized.

3.1.1 BROMINATED FLAME RETARDANTS (BFRS)

Flame Retardants (FRs) are a group of chemicals used to reduce the flammability of combustible materials such as plastics, roots or textiles. The most abundantly used FRs contain bromine and compounds of this family are known as BFRs. They are added to a wide variety of consumer goods, including electronics, furniture, building materials, and automobiles, to make them less flammable. Depending on their mode of incorporation into the polymers, BFRs can be classified as additive (the most frequently detected in environment due to their potential to leak from treated consumer products), reactive, or polymeric.

The most investigated additive BFRs are Polybrominated diphenyl ethers (PBDEs), polybrominated biphenyls (PBBs) and Hexabromocyclododecane (HBCDs). Each class may include multiple congeners (chemical substances with similar structure, origin or function) and their chemical structure and the main physicochemical properties of these compounds are presented in Figure I.16 and Table I.2.



a. PBDEs (209 congeners) b. PBBs (209 congeners) c. HBCDs (3 congeners)

Figure I.16. **Chemical structures of major BFRs compounds**

Table I.2. **Physicochemical properties of PBBs, PBDEs, and HBCDs**
Adopted from The Handbook of Environmental Chemistry[60]

Chemical	Acronym	Formula	Molecular Mass	Melting point (°C)	Decomposition point (°C)	Solubility H_2O (µg/L25°C)	Log K_{ow}
PBBs	beta-BB	$C_{12}H_4Br$	627.4	124–248	300–400	11	7.20
	octa-BB	$C_{12}H_2Br_8$	785.2	200–250	435	30–40	5.53
	nona-BB	$C_{12}HBr_9$	864.1	220–290	435	Insoluble	
	deca-BB	$C_{12}Br1_0$	943.0	380–386	395 > 400	<30	8.58
PBDEs	tetra-BDE	$C1_2H_6Br_4O$	485.8	82.3	-	4.7	5.87–6.16
	penta-BDE	$C_{12}H_5Br_5O$	564.7	81.0	>200	4.4	6.64–6.97
	octa-BDE	$C_{12}H_2Br_8O$	801.5	200	-	-	8.35–8.90
	deca-BDE	$C_{12}Br_{10}O$	959.2	290–306	>320	20–30	9.97
HBCD	α-HBCD	$C_{12}H_{18}Br_6$	641.7	179–181	>190	48.8	5.07
	β-HBCD			170–172		14.7	5.12
	γ-HBCD			207–209		2.1	5.47

Source of human exposure

PBDEs can be found in plastics, textiles, electronic castings and circuitry; HBCDs in thermal insulation in the building industry while PBBs are used in consumer appliances, textiles and plastic foams (EFSA). BFRs have the tendency to be extremely stable and persistent in the environment, having long half-lives in soils, sediments, air, or biota[61]. Because of their tendency to accumulate in living organisms, these chemicals are detected in foods, mainly fish, but also meat and dairy products.

The potential for organic compounds to bioaccumulate and widespread in the environment is a direct consequence of their physicochemical properties such as lipophilicity and resistance to degradation. One way to obtain an estimate of the human exposure to environmental contaminants is through biomarkers and specifically by measuring the presence of chemical compounds in storage tissues (adipose tissue, hair, nails) in blood (i.e. levels in plasma and serum) and in excreted liquids (i.e. urine and breast milk).

BFRs are known to be extremely lipophile, this degree of bioaccumulation depending on a number of parameters including their molecular weight and octanol-water partition coefficient (Log K_{OW}) which represents a measure of the tendency of a compound to move from the aqueous phase into lipids. The half-life of BFRs appears to be related to the number of bromine atoms per molecule. For instance, the average half-life of BDE-47, BDE-99 and BDE-153 are respectively 1.8 years (1.4 - 2.4), 2.9 years (1.8 - 4.0) and 6.5 years (3.6 - 12,4)[62]. Authors also reported half-life of 64 days (range 22-210 days) for HBCDs.

Being excreted in breast milk, BFRs represent a significant exposure for infants and small children and may have a significant impact on their health.

Children, as well as adults are also mainly exposed through indoor air inhalation and dermal contact but it has been reported that dust ingestion was the dominant exposure pathway for most studied BFRs (compared to indoor inhalation and dermal contact), especially for infants and toddlers who have higher exposures than older children[63]. In the same study, findings reveal that the highest indoor house dust concentrations of PBDEs are found in North America and for BDE-209 in Europe and China (Figure I.17).

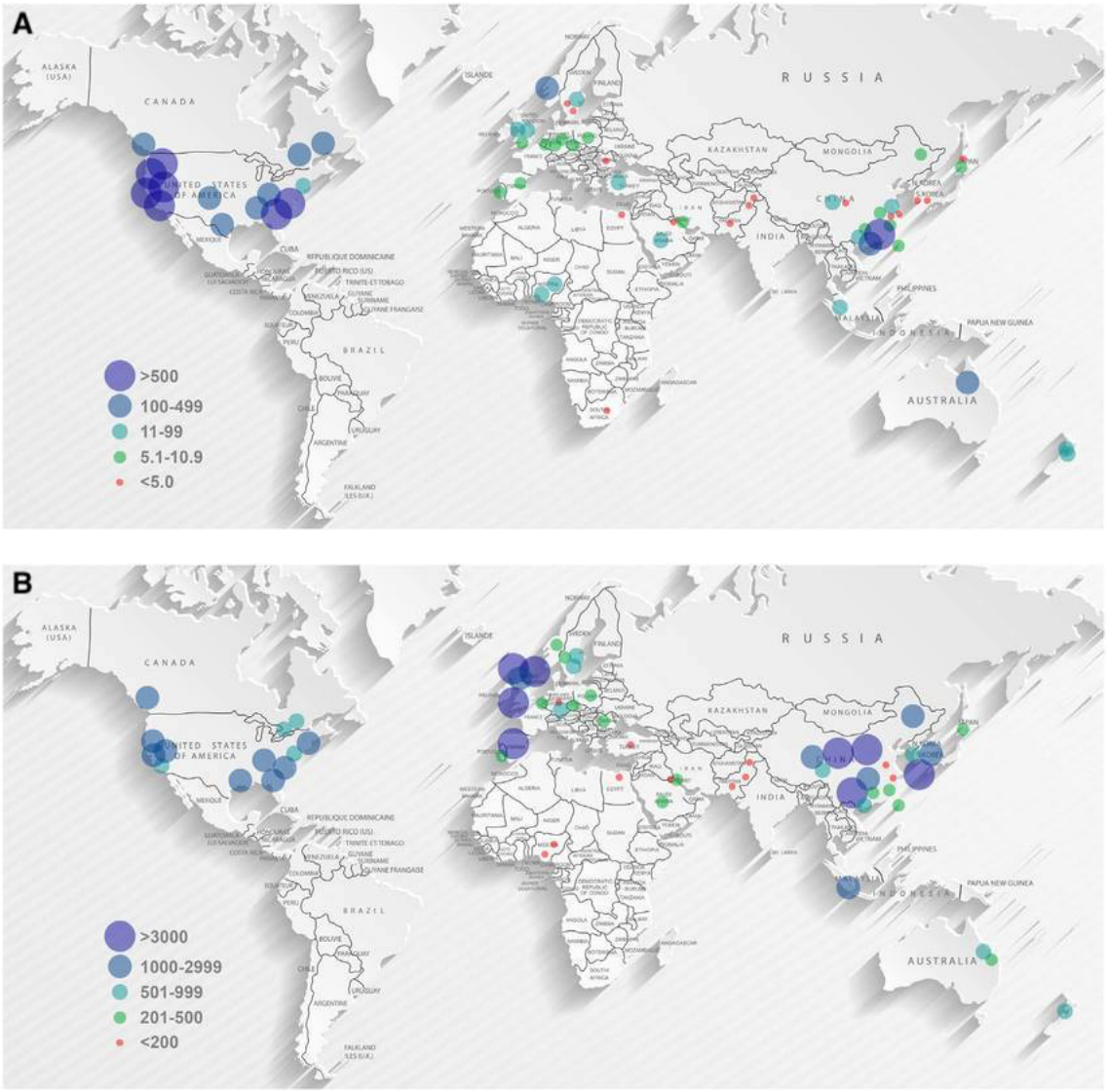


Figure I.17. **Worldwide distribution of median PBDEs congeners indoor house dust concentrations** A) BDE-47 (ng/g). B) BDE-209 (ng/g).
Adapted from Malliari and Kantzi[63]

Effects on human health

In terms of toxicity, particularly neurotoxicity, most studies have been conducted using animal models such as mice or zebrafish. Mice exposed on postnatal day (PND) 10 (i.e. the peak of the brain growth spurt) to PBDEs or HBCDs develop permanent aberrations in spontaneous behavior and habituation (decrement in response as a result of repeated stimulation not due to peripheral process like receptor adaptation or muscular fatigue) capability, and changes in the development of neuromotor systems[64,65].

In zebrafish, it has been shown that BDE-209 congener affects expression of neurological pathways and alters the behavior of larvae, whereas parental chronic low dose exposure affects growth and reproduction and elicits neurobehavioral alterations in offspring[66]. The exposure to BDE-47 and its metabolite 6-OH-BDE-47 also affects the locomotion behavior of both larval and juvenile zebrafish[67]. Several studies about the effects on reproduction have also been conducted using animal models. Pregnant rats were exposed to BDE-47 from gestation day 8 until PND 21 and male reproductive outcomes were analyzed on PND 120 in offspring[68]. Exposed animals had significantly smaller testes, displayed decreased sperm production per testis weight, had significantly increased percentage of morphologically abnormal spermatozoa, and showed an increase in spermatozoa head size. Also, perinatal BDE-47 exposure led to significant changes in testes transcriptome, including suppression of genes essential for spermatogenesis and activation of immune response genes.

Even if BFRs are excreted through breast milk and that therefore breastfeed infants are exposed to BFRs, the epidemiological evidence that exposure to human milk containing background levels of such chemicals would pose a serious health hazard is limited and insufficient[69]. One study reported a correlation between infant weight at birth and length at birth with the levels of PBDEs congeners (47, 99, 100 and 153) in Northern Tanzania[70]. Another study conducted in China in term of occurrence and temporal trends showed that daily dietary BFRs intake for nursing infants is much higher than that for adults[71]. As for the assessment of the potential effects on health, the current scientific literature is contradicting. For example, in the same study, the risk assessment evaluated using the Margin Of Exposure (MOE) approach (a tool used by risk assessors to consider possible safety concerns arising from the presence in food and feed of substances which are both genotoxic _they may damage DNA_ and carcinogenic) concluded that dietary BFRs intake for nursing infants was unlikely to pose significant health risks while a study of BFRs in placental tissues suggest a potential alteration of thyroid hormone function[72].

Additionally, as conducted by Leonetti and colleagues[72], most studies related to health issues in association with PBDEs are related to a possible disruption of thyroid hormones[73–75], mainly due to the

similarity in chemical structures of PBDEs and thyroid hormones triiodothyronine (T3) and thyroxin (T4), and thus the potential for PBDEs to mimic and disrupt homeostatic conditions[60].

Finally, recent studies have suggested that BFRs could play a role in the epidemic of type 2 diabetes (T2D). A study using the E3N prospective cohort of French women was conducted to evaluate the association between dietary exposure to BFRs and T2D risk. Findings suggest an association (positive linear trend) between dietary exposure to HBCDs and T2D risk starting from the 2nd quintile group (HR: 1.18; 95% CI: 1.06–1.30) to the 5th quintile group (HR: 1.47; 95% CI: 1.29-1.67) when compared to the 1st quintile group. Authors also found positive although non-linear associations between dietary exposure to PBDE and T2D risk, with an increased HR only for the 2nd and 4th vs. 1st quintile groups (HR: 1.12; 95% CI: 1.02–1.24, and HR: 1.20; 95% CI: 1.08–1.34, respectively)[76].

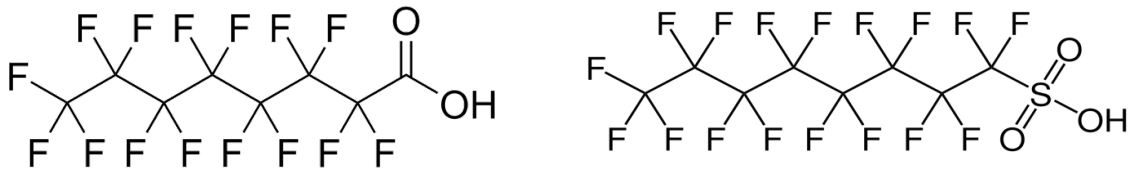
Because of the threat POPs, including BFRs, may pose to human health and the environment, such substances are regulated under the Stockholm Convention that was adopted in 2001 including 152 signatories and 183 parties. The effectiveness of this Convention, whose broad aim is to protect human health and the environment by controlling the releases of POPs, has been evaluated in several studies. A time series analysis of atmospheric POP concentrations from 15 monitoring stations in North America and Europe concluded that a decade of air monitoring data has not been sufficient for detecting general and statistically significant effects of the Stockholm Convention[77].

Results suggest that the observed changes are the result of national regulations enforced prior to the implementation of the Stockholm Convention, rather than to the enforcement of the provisions laid out in the Convention. Other studies on BFRs showed a decrease in the detected levels that may be associated with the implementation of the Stockholm Convention. For example, a Californian study published in 2015 found significant declines of some PBDEs congeners levels in breast milk between 2003-2005 and 2009-2012 (from 67.8ng/g lipid to 41.5ng/g lipid)[78]. Another study conducted in China with -47, -99 and -100 congeners showed significant relative decreases in the human milk levels with an average of 45%, 48%, and 46% decrease from 2007 to 2011, for the three congeners respectively[79]

3.1.2 PER- AND POLYFLUORINATED ALKYLATED SUBSTANCES (PFASS)

Per- and polyfluoroalkylated substances (PFASs) are a vast group of chemicals widely found in a large range of products used by consumers and industry. Most of them are impermeable to grease, water and oil. For this reason, they are used for many different applications including in stain- and water-resistant fabrics and carpeting, cleaning products, paints and fire-fighting foams, as well as in limited, authorized uses in cookware and food packaging and processing (U.S Food and Drug Administration).

Among all PFASs, the perfluorooctanoic acid (PFOA) and the perfluorooctanesulfonic acid, also known as perfluorooctanesulfonate (PFOS), have been the most widely used and are therefore the object of monitoring and research on their effects on human health and the environment. PFOA and PFOS are very persistent in the environment and in the human body and there is evidence that exposure to such substances can lead to adverse human health effects. Tolerable weekly intakes of PFOA and PFOS set up to 6 $ng \cdot kg^{-1} \cdot bw \cdot week^{-1}$ (based on the daily calculated intakes resulting in a critical serum concentrations and outcomes, the weight and the half-life of the contaminant[80]) and 13 $ng \cdot kg^{-1} \cdot bw \cdot week^{-1}$, respectively (EFSA). The chemical structure and the main physicochemical properties of these compounds are described in Figure I.18 and Table I.3.



a. PFOA b. PFOS

Figure I.18. **Chemical structures of major PFASs compounds**

Table I.3. **Physicochemical properties of PFOA and PFOS**

Chemical	Formula	Molecular Mass	Melting point (°C)	Decomposition point (°C)	Solubility H_2O (g/L)	Log K_{ow}
PFOA	$C_7HF_{15}O$	414.07	55–56	-	3.4	4.59
PFOS	$C_8F_{17}SO_3H$	500.1	> 400	-	0.57	5.26

Source of human exposure

People can be exposed to PFASs through various ways, notably food that may be contaminated by contaminated soil and water used to grow the food or from food packaging. The widespread use of PFASs and their ability to remain intact in the environment mean that over time PFASs levels from past and current uses can result in increasing levels of environmental contamination. (Figure I.19).

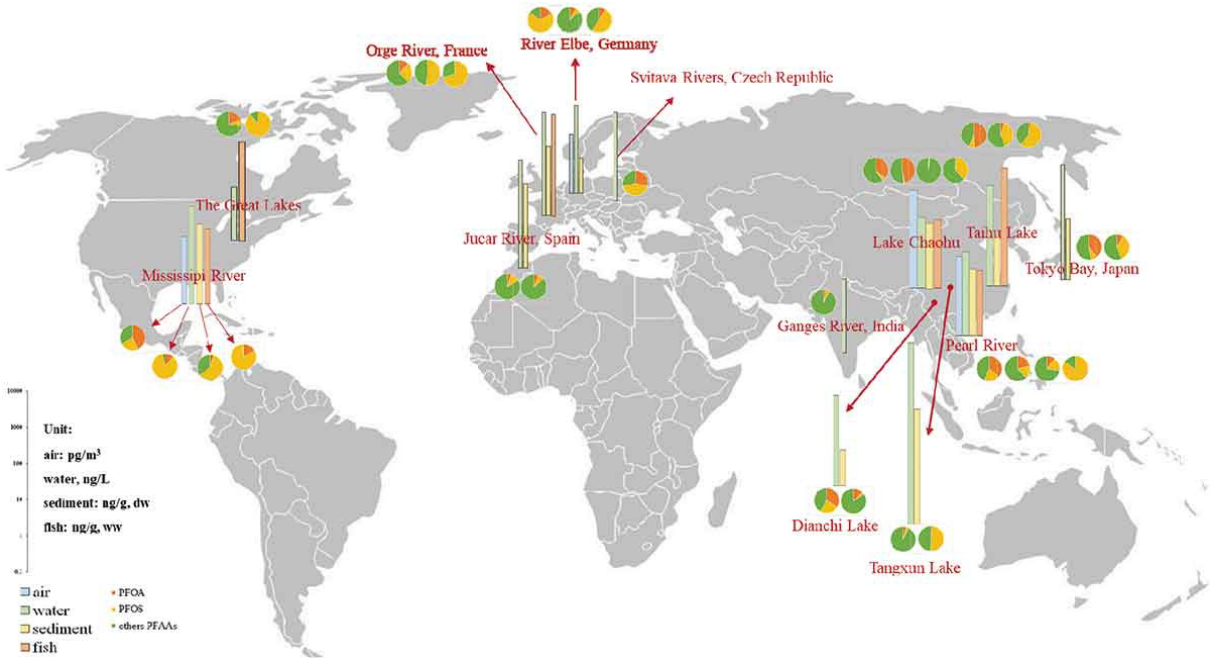


Figure I.19. **The occurrence of perfluoroalkyl acids in the global environment (including air, water, sediment and fish)**
Adapted from Liu and colleagues[81]

In France, for example, Bach and colleagues[82] performed a study that estimated the extent of contamination with PFASs of the river Orge. They estimated that 4295 kg of PFHxA, 1487 kg of 6:2FTSA, 965 kg of PFNA, 307 kg of PFUnDA, and 14 kg of PFOA were discharged in the river by two facilities in 2013. It was found that chlorination (a method of water treatment) had no removal efficiency and even if the total PFASs concentrations were high in the treated water, ranging from 86 to 169 ng/L, they did not exceed the currently available guideline values.

Workers exposed professionally to PFASs have higher levels of PFASs exposure than a non-occupationally exposed group[83]. In a retrospective U.S study of an aging population, findings showed that participants with high cumulative workplace exposure (work in occupations and industries known to use PFASs) had 34% higher serum PFOS levels compared to participants without occupational exposure, adjusted for age, sex and income and serum PFOS levels were 26% higher for participants with longer occupational exposure durations[84].

To determine whether bladder cancer is associated with exposure to (PFOS) in an occupational cohort, a study among former employees of a facility of PFOS production was conducted[85]. Eleven cases of primary bladder cancer were identified from the surveys and compared with employees in the lowest cumulative exposure category, the relative risk of bladder cancer was 0.83 (95% CI = 0.15–4.65), 1.92 (95% CI = 0.30–12.06), and 1.52 (95% CI = 0.21–10.99) with a cumulative exposure of 1, 1–5, 5–10, and >10 years.

As for BFRs, PFASs can also be found in blood and breast milk with known adverse effects of prenatal exposure to PFASs in developmental outcomes in offspring[86,87]. In the meantime, significant correlation was found between the parity of mothers and PFASs concentrations in human milk and it was reported that primiparas showed higher PFASs levels in human milk than multiparas in France, Italy, and Belgium[88].

In contrast to BFRs and most other POPs, they do not tend to accumulate in fat tissues but bind to serum albumin and other cytosolic proteins and accumulate mainly in the liver, the kidneys, and bile secretion[83]. They are considered as amphiphilic (molecules having a polar water-soluble group attached to a water-insoluble hydrocarbon chain) compounds[84] and their half-life in human serum was respectively set 5.4 and 3.8 years for PFOS and PFOA in 2007[89] while findings from a more recent study (2018) indicates a decrease from 3.4 and 2.7 years respectively[90].

Effects on human health

PFOS and PFOA have been associated with liver enlargement in rodents and nonhuman primates in addition to hepatocellular adenomas in rats and a number of short-term studies in rats and mice have shown that PFOS and PFOA are capable of inducing peroxisome (organelle involved in catabolism of very long chain fatty acids) proliferation through the activation of PPAR-α (peroxisome proliferator–activated receptor-alpha) known to be involved in tumour (primarily liver) induction by a number of nongenotoxic carcinogens in the rodents[91].

In term of reproduction, a study reveals that zebrafish embryos exposed to 16 µM PFOS during a sensitive window of 48-96 hour post-fertilization (HPF) disrupted larval morphology at 120 HPF and malformed zebrafish larvae were characterized by uninflated swim bladder, less developed gut, and curved spine[92]. Additionally, whole genome microarray was used to identify the early transcripts dysregulated following PFOS exposure and a total of 1278 transcripts were significantly misexpressed ($p<0.05$) while 211 genes were changed at least two-fold upon PFOS exposure in comparison to the vehicle-exposed control group. Chronic exposition to PFOS have also been reported to reduce sperm quality and expression of key genes involved in hormone pathways[93].

Due to their persistence, as well as ubiquity in the environment caused by long-range transport, current evidence suggests that the bioaccumulation of certain PFASs may cause serious health conditions in humans.

Recently, in a case-control study nested in the French E3N cohort PFASs (PFOA and PFOS) circulating levels were differentially associated with breast cancer risk[94]. Findings showed a positive linear associations between PFOS concentrations and the risk of ER+ (3rd quartile: OR = 2.22 [CI = 1.05–4.69]; 4th quartile: OR = 2.33 [CI = 1.11–4.90]) and PR+ tumours (3rd quartile: OR = 2.47 [CI = 1.07–5.65]; 4th quartile: OR = 2.76 [CI = 1.21–6.30]). When considering receptor-negative tumours, only the 2nd quartile of PFOS was associated with risk (ER−: OR = 15.40 [CI = 1.84–129.19]; PR−: OR = 3.47 [CI = 1.29–9.15]). While there was no association between PFOA and receptor-positive BC risk, the 2nd quartile of PFOA was positively associated with the risk of receptor-negative tumours (ER−: OR = 7.73 [CI = 1.46–41.08]; PR −: OR = 3.44 [CI = 1.30–9.10]).

Earlier in 2017, in a case control study of Inuit women from Greenland, significant, positive associations between breast cancer risk and both of them with other classes of PFASs (PFHpA, PFDA,PFUnA,PFDoA) were also observed[95] while in the California Teacher Study, a similar retrospective case-control study in which PFASs levels for cases were measured after diagnosis[96]. Overall, these results are limited but suggestive that exposure to PFASs may increase breast cancer risk though further studies are necessary to strengthen the evidence.

The epidemiological evidence on PFASs exposure as a risk factor for diabetes is limited and inconsistent although the availability of supporting data and studies. Regarding T2D, a prospective cohort study identified an association between PFOA with incident diabetes and microvascular disease and the results suggest that exercise and diet may attenuate the diabetogenic association of PFASs[97]. Some of them report positive associations[98,99] while others report inverse[100] or null associations[101].

3.2 PERSISTENT ORGANIC POLLUTANTS AND DNA METHYLATION

For the purpose of this section, the term "POPs" will refer not only to BFRs and PFASs but also to other pollutants. We are interested in studies focusing on DNA methylation.

Effect of POPs on DNA methylation is not completely established even if alterations of epigenetics mechanisms are known to be linked to environmental exposures with adverse health effects. Also, most of published studies were focused on prenatal and early-life exposures which can be explained by the fact that the epigenome undergoes extensive reprogramming throughout fetal development at gametogenesis and early embryo preimplantation, representing vulnerable stages to environmental exposure[102] (Figure I.20). Additionally, POPs can cross the placenta and reach the newborn through breast milk. Generally, in these studies, only global methylation is evaluated.

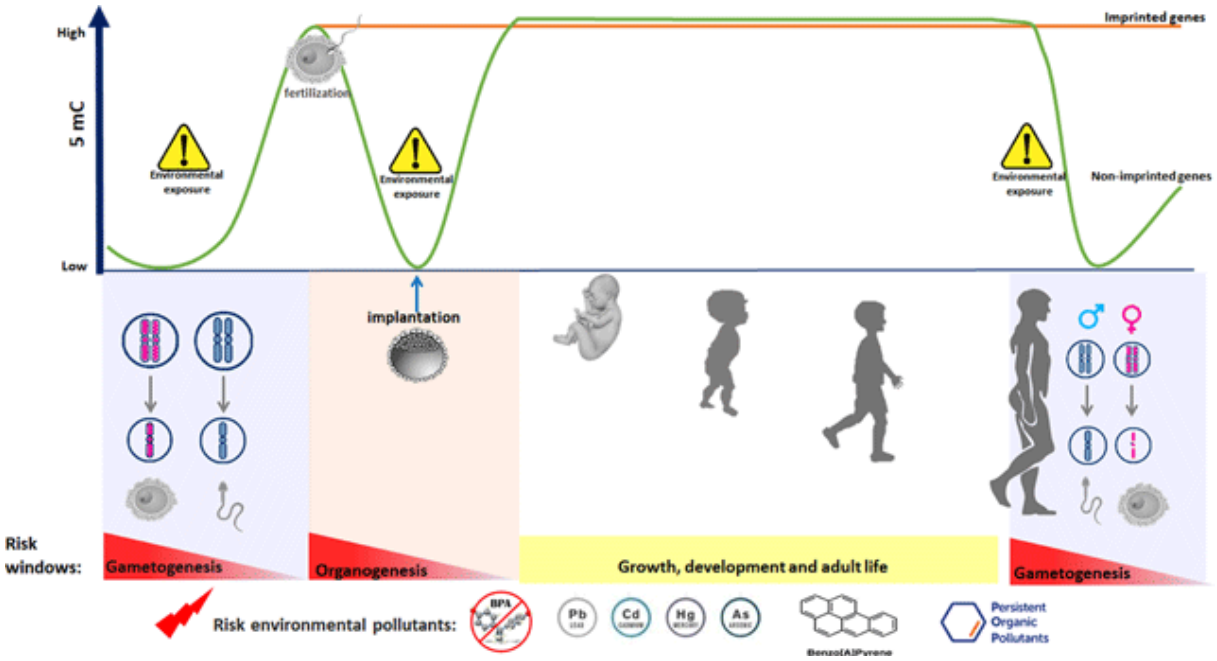


Figure I.20. **Susceptibility windows of DNA-methylation due to environmental pollutants**
Adapted from Alvarado-Cruz and colleagues[102]

As previously reported in the section related to DNA methylation, Alu and LINE-1 elements are widely used as markers of global methylation. Alu elements (repetitive elements that comprise approximatively 10 % of the human genome), have wide-ranging influences on gene expression and contribute to genome evolution and gene regulation[103]. They belong to a class of retroelements termed SINEs (Short INterspersed elements) and are primate specific. These elements are non-autonomous, in that they

acquire *trans*-acting factors for their amplification from the only active family of autonomous human retroelements: LINE-1 that represents around 20 % of the human genome.

In a birth cohort from Mexico, findings suggested that co-effect of DDT (dichlorobiphenyl trichloroethane) and PBDEs exposure induce global hypomethylation[104]. This result was confirmed in another independent study[105].

Regarding PFASs, a study of 363 mother-infants suggested that prenatal PFOS exposure may be associated to Alu DNA hypomethylation in cord blood[106] while another study from a US-based population found that in utero PFOA exposures also induce global hypomethylation in cord blood[107]. On the other hand, using Luminometric Methylation Assay (LUMA), which is a method that allows to capture DNA methylation using restriction enzymes and Pyrosequencing[108], no association was found between DNA methylation and BDE-47 congener. However, in the same study, global hypermethylation was found to be associated with high serum levels of some POPs in contradiction to a previously mentioned study and others that used different design.

A study conducted within the British Birth Cohort examined association between BDE-47 congener from maternal blood and methylation Tumour Necrosis Factor alpha (TNFα) promoter in cord blood. TNFα is a cytokine that plays important roles in inflammation and metabolism mechanisms. Results showed that a decrease of *TNFα* methylation is associated with an increase in TNFα protein level in cord blood and provided evidence that *in utero* exposure to PBDEs may epigenetically reprogram the offspring's immunological response through promoter methylation of a proinflammatory gene[109]. Finally, some studies suggest that POPs are potential germline epimutagens and could be tied to preconception exposure[110–112].

4. SUMMARY AND OBJECTIVES

Mutational signatures

Mutational signatures refer to patterns in the occurrence of somatic mutations that might be uniquely ascribed to particular mutational process. Tumours mutation catalogues can reveal mutational signatures but are often consistent with the mutation spectra produced by a variety of mutagens. To date, after the analysis of tens of thousands of exomes and genomes from about 40 different cancer types, tens of mutational signatures characterized by a unique probability profile across the 96 trinucleotide-based mutation types have been identified, validated and catalogued.

After the introduction of the original framework for the formal definition and analysis of mutational signatures, several other mathematical methods and computational tools have been proposed to detect mutational signatures and estimate their contribution to a given catalogue as well as their potential association with an endogenous or exogeneous exposures.

In termed of association between mutational signatures and environmental exposures, most findings were mainly related to UV light, tobacco consumption or aristolochic acid.

Epigenetic signatures of Persistent Organic Pollutants

Epigenetics is defined as the study of mitotically heritable yet potentially reversible, molecular modifications to DNA and chromatin without alteration to the underlying DNA sequence. DNA methylation, one of the most studied epigenetics marks is known to be dynamic in response to environmental stimuli and have been associated with a wide range of environmental exposure and multifactorial disease.

POPs are organic compounds that are widespread in the environment. Because of their persistence, they are able to bioaccumulate with major impacts on human health.

Regarding's epigenetic signatures and particularly DNA methylation, and with regards to the existing literature that supports the role of POPs-associated methylation as a potential mediator of POP-associated health effects in humans, more research is required as most of conducted studies were focused on LINE-1 or Alu elements as marks of global methylation.

Objectives and results

This thesis has two main objectives:

1. Mutational signatures: review contributions to epidemiology and evaluate existing methods
   - ➔ We review the existing literature related to mutational signatures linked to environmental exposures and lifestyle and their implication in the development of lung adenocarcinoma (**Papers 1 and 2,** *published*).
   - ➔ We introduce a probabilistic model for simulating mutational signatures and catalogues and conduct an original empirical comparison of the performance of developed tools for mutational signatures analysis (**Paper 3,** *published*).
2. Epigenetic signatures of POPs: study of the association between two important families of EDCs and DNA methylation using the French prospective E3N cohort
   - ➔ We evaluate the association between BFRs and DNA methylation (**Paper 4,** *submission in progress*).
   - ➔ We evaluate the association between PFASs and DNA methylation (**Paper 5,** *submission in progress*)

CHAPTER II:

ENVIRONMENT AND LIFESTYLE INFLUENCE ON MOLECULAR FEATURES

In this chapter, we describe how recent advances in the study of mutational and epigenetic signatures in tumours provide new opportunities to understand the role of the environment and lifestyle in cancer development. In the first part of the chapter, that is the object of our recent publication in the journal Current Opinions in Oncology[113], we discuss how such recent advances in the study of mutational and epigenetic signatures may be applied to the study of the etiology of cancer and we provide some interesting examples. In the second part of the chapter, that has been presented in a separate publication that has attracted media coverage (https://www.inserm.fr/actualites-et-evenements/actualites/non-cancer-est-pas-principalement-hasard), we extend the application of mutational signatures to contribute to the debate around the "bad luck" hypothesis related to cancer development (incorrectly popularized as "2/3 of cancers *are due* to errors in DNA replication during cell division and therefore to *intrinsic* and unpreventable causes"). In such work we introduce an analysis showing that smoking-induced mutations are more predictive of cancer risk than the lifetime number of stem cell divisions.

Contribution
Co-author, contributed to the review and the figures, read and approved the final reports.

1. Environmental exposures associated mutational and epignetics signatures 71
1.1The exogeneous causes of mutational signatures 71
1.2 Exposures related epigenetics signatures in tumour tissue 74
2. Exposure to smoking, lung adenocarcinoma development and the "bad" luck cancer theory 76
2.1 The "bad luck" debate: stem cell divisions, driver mutations and cancer risk 77
2.2 Predicting lung cancer risk via *extrinsinc* mutations 80
3. Conclusion 82

1. ENVIRONMENTAL EXPOSURES ASSOCIATED MUTATIONAL AND EPIGNETICS SIGNATURES

Cancer-related mutational events have been investigated for decades and, in more recent years, numerous epigenetic hallmarks of cancer have been identified but only with the recent development of high throughout sequencing and the resulting wider availability of genomic sequences and epigenomic data from thousands of cancer exomes and genomes have made possible to identify numerous distinct mutational and epigenetic signatures some of which have been associated to environmental exposures, carcinogens and factors related to lifestyle.

1.1THE EXOGENEOUS CAUSES OF MUTATIONAL SIGNATURES

The idea that carcinogens leave fingerprints is not novel[114]. The notion that exposure to ultraviolet radiation (UV) caused predominantly the transition cytosine to thymine (C > T) and tobacco smoke predominantly caused the transversion cytosine to adenine (C > A) has been established experimentally several decades ago[115], well before the development of sequencing technologies. However, the generation of a large number of tumour sequences (cancer exomes or whole genomes) and the development of appropriate mathematical methods greatly improved the capacity to identify such fingerprints[116]. While initially some of the mutational signatures have been linked to specific factors only on the basis of biological prior knowledge of their mutational effects[117], more recently experimental studies and studies that coupled individual information about environmental exposures and lifestyle with tumour sequencing data are providing useful information to establish the causes of some signatures. In the following paragraphs of this section, we review some examples of exposures proposed as the origin of specific mutational signatures.

1.1.1 TOBACCO

To investigate mutational signatures in tobacco-related cancers, Alexandrov and colleagues studied the cancer genomes from 2 490 smokers and 1 063 never smokers[118]. For each cancer, they extracted a list of mutational signatures and estimated their contributions to the complete mutational catalogue. By comparing the mutational signatures identified in cancer genomes in smokers and non-smokers, they found that signatures 2,4,5,13, and 16 in COSMIC were more prevalent in smokers than in non-smokers. Signature 4, for example, appears to be a strong signature related to exposure to tobacco smoke as it is observed in tumours strongly associated with tobacco smoking (e.g. lung squamous cell carcinomas, lung adenocarcinomas, larynx and liver cancers) and its prevalence is higher in smokers than in non-smokers. Signature 4 was associated with pack-years smoked and it was not found in tumour tissues from organs not directly exposed to tobacco smoke. Notably, this signature is mostly characterized by C > A transversions, an observation consistent with previous knowledge about the mutagenic effects of

tobacco smoke, and its mutation profile is very close to that caused by exposure to some chemicals present in tobacco smoke such as benzo[a]pyrene that earlier experimental studies have demonstrated to be a carcinogen[119].

1.1.2 AFLATOXIN B1

Another interesting example of exposure linked to specific mutational signatures is exposure to aflatoxin B1 (AFB1), a common contaminant in a variety of foods such as peanuts, corn and grains that represents a major public health problem in some regions of Africa and Asia as it strongly increases the risk of hepatocellular carcinoma (HCC), especially when associated with hepatitis B. An interesting study that investigated mutational signatures in human cell lines and liver cancers in mice exposed to AFB1and corroborated the results with analyses of signatures extracted from human HCC genomes from a geographical region in which exposure to AFB1 is well documented, provided strong support to the likely link between exposure to AFB1 and signature 24[120]. Such signature has been found only in the genome of HCCs.

1.1.3 IONIZING RADIATION

The tumourigenic effect of ionizing radiation particularly in the context of the iatrogenic effects of cancer treatment is also an interesting application of mutational signatures. Analyses of the genome of 12 second malignancies associated with radiation treatment of primary tumours identified two genomic imprints or signatures not present in cancers not exposed to ionizing radiation[121]. These signatures, being characterized by small deletions occurring with similar density across the genome as well as by balanced inversions, are not captured by the common methods to extract mutational signatures based on base substitutions. To overcome the scarcity of genomic sequences for radiotherapy-induced cancers, it was proposed to conduct combined analyses of mutational catalogues from ionizing radiation-induced cancers in human tumour sequences and in tumour sequences from mice models[122]. This type of analysis identified two signatures linked to ionizing radiation that had not been previously identified and may represent a useful approach also for other exposures.

1.1.4 UV LIGHT

The typical C > T transitions induced by exposing experimental systems to UV light, are characteristic of signature 7 that is found in melanomas and head and neck cancers. These observations have led to propose UV light as the cause of signature 7[117].

1.1.5 ARISTOLOCHIC ACID

Aristolochic Acid (AA) is a natural compound contained in plants from the *Aristolochiaceae* family used in some herbal remedies or traditional medicines. AA is a known nephrotoxic phytochemical causing endemic nephropathy and a carcinogen that was previously associated with urothelial cancers of the upper urinary tract. A study based on urothelial tumours from 15 patients with endemic nephropathy identified signature 22 and linked it to AA exposure[123]. An important aspect of this study is that it demonstrates that such signature can be observed with exome sequencing of DNA from formalin-fixed paraffin-embedded tumour samples even at low sequencing coverage (less than 10X). Signature 22 is mostly characterized by A > T or T > A transversions that were found in experimental studies based on human renal cells exposed to AA[124] and in a series of urothelial cancers in patients with a documented exposure to AA[125]. Evidence of exposure to AA was found in the genomes of a minority of bladder cancers (4 out of 110 tumour samples) from Singapore and China[126] and, interestingly, in 11 of 93 HCCs, a type of cancer not known to be associated with exposure to AA[124]. The presence of the AA-related signature was found also in clear cell renal cell carcinomas[127,128]; with a particularly high prevalence in cases from regions in Romania where Balkan Endemic Nephropathy is prevalent and due to widespread exposure to AA[129]. These studies do not refer explicitly to specific COSMIC signatures, but their results are consistent with the proposed link between COSMIC signature 22 and exposure to AA.

1.2 EXPOSURES RELATED EPIGENETICS SIGNATURES IN TUMOUR TISSUE

As previously described in chapter I, DNA methylation is an epigenetic mechanism consisting in the addition of a methyl group to the cytosine base of the CpG nucleotides of the DNA sequence. DNA methylation modulates gene expression by influencing DNA transcription and it is involved in many biological processes, including the response of cells to external stress. Modifications of physiologic DNA methylation patterns are associated with the development of many diseases, including cancer for which altered DNA methylation has been observed in early stages of carcinogenesis and for many cancer types[130]. Features common to many cancer tissues are global hypomethylation, which causes genome instability[131], and hypo- or hypermethylation of specific loci, causing overexpression of oncogenes and under expression of tumour suppression genes.

Many studies have been conducted to identify methylation signatures of risk that may be used for primary prevention or methylation markers to detect cancer in early stages and contribute to secondary prevention. Such efforts have been supported by the increasing availability of a variety of molecular techniques able to profile whole genome methylation or identify differentially methylated regions[132]. As far as methylation markers of risk are concerned, of particular interest are the studies that established a relationship between some environmental and lifestyle factors and in particular cigarette smoking and the levels of methylation in DNA from blood. The methylation levels of thousands of CpG sites have been found to be altered in smokers compared with non-smokers and such alterations appear to be associated with smoking duration and intensity[133,134]. There is strong evidence from analyses of tobacco-related alterations of methylation of blood DNA from former smokers that for some CpGs methylation levels reverse in a few years after quitting smoking to the levels observed in non-smokers while for other CpGs the alterations are observed even decades after quitting smoking.

The study conducted by Alexandrov and colleagues that scrutinized tobacco-related mutational signatures in 5 243 tobacco-related cancers, also analyzed methylation profiles of tumours to assess the presence of the tobacco-related methylation signatures that have been identified in DNA from blood[118]. Average differences in DNA methylation larger than 5% between smokers and lifelong-nonsmokers were observed in tumour tissue of lung adenocarcinomas cases and oral cancer cases, but not in tumour tissues of other smoking-related cancer types. The main differences were observed for lung adenocarcinomas where in smokers 369 CpGs were hypomethylated and 65 hypermethylated; for oral cancer only 8 differentially methylated CpGs were observed, 5 of whom were hypomethylated. Interestingly, none of these CpGs are among those found to be differentially methylated in blood or buccal cells of smokers and non-smokers.

In another study a tobacco-related methylation index was estimated in cancer and surrounding normal tissue of various cancer types including lung cancer. The DNA methylation-based index associated with

exposure to cigarette smoking was developed from 1 501 differentially methylated CpGs in DNA from epithelial buccal cells of smokers and non-smokers[135]. The methylation index was then calculated using methylome data separately for normal and cancer tissue and it was found to be extremely accurate in discriminating between normal and cancer tissue for lung cancer and other cancer types; the index was also able to discriminate between lung lesions that regressed from those that progressed.

Stueve and colleagues searched for methylation signatures associated with tobacco smoke in normal tissue surrounding tumour tissue in 237 lung cancer cases using methylation data generated with the Infinium HumanMethylation450 Bead Chip array and identified 7 CpGs in which hypomethylation was associated with cigarette smoking[136]. For all these CpGs the association between hypomethylation and cigarette smoking was confirmed with TCGA methylation data. Five of the 7 CpGs corresponded to CpGs for which tobacco-related hypomethylation had been previously observed in DNA from peripheral blood. Notably, for one the 7 CpGs (i.e. cg05575921) an association between hypomethylation and lung cancer risk independent of the exposure to tobacco smoke had been previously reported[53,54].

In an analysis using a line of epithelial cells exposed to cigarette smoke condensate (CSC) aimed at understanding the possible functional consequences of hypomethylation at the identified CpGs, induced gene expression was evaluated in the 1Mb window flanking the CpGs. Hypomethylation levels in four CpGs were associated with induced expression of the genes *AHRR, CYP1B1*, *ENTPD2* in the CSC exposed cell line. Such observation, confirmed in the TCGA data from lung cancer, is particularly interesting as in the promoters of the *AHRR, CYP1B1*, and *ENTPD2* genes are present binding sites for the aryl hydrocarbon receptor (AHR), a transcription factor involved in detoxification and bioactivation of pro-carcinogens in tobacco smoke, suggesting a possible pathway linking smoking induced methylation to lung cancer. Interestingly, in addition to the observed association with tobacco-induced hypomethylation at specific loci, Stueve and colleagues noticed that increased expression of the *AHRR* and, to a lesser extent, *CYP1B1* genes was also associated with the tobacco-related C > A substitutions[25].

The debate about the interpretation of the associations between cigarette smoking, alterations of DNA methylation and lung cancer risk, is still open as results from a recent Mendelian randomization study would not be consistent with the hypothesis of a causal link between the tobacco-related alterations in methylation levels and lung cancer risk[137].

2. EXPOSURE TO SMOKING, LUNG ADENOCARCINOMA DEVELOPMENT AND THE "BAD" LUCK CANCER THEORY

Lung cancer is the third most common cancer worldwide and it is well-established that tobacco smoke is the main cause. Smoking is also a major cause of other cancers such as cancers of the bladder, oral and nasal cavity, oropharynx, larynx, kidney, bowel, oropharynx, stomach, liver, esophagus and pancreas[138].

In 2017, it was estimated that over 90% of lung cancer cases among men and over 80% of cases among women worldwide are attributable to tobacco use (WCRF). A study conducted in France in 2015 attributed 20% of new cancers cases (68 680) to tobacco consumption (Figure II.1)[139].

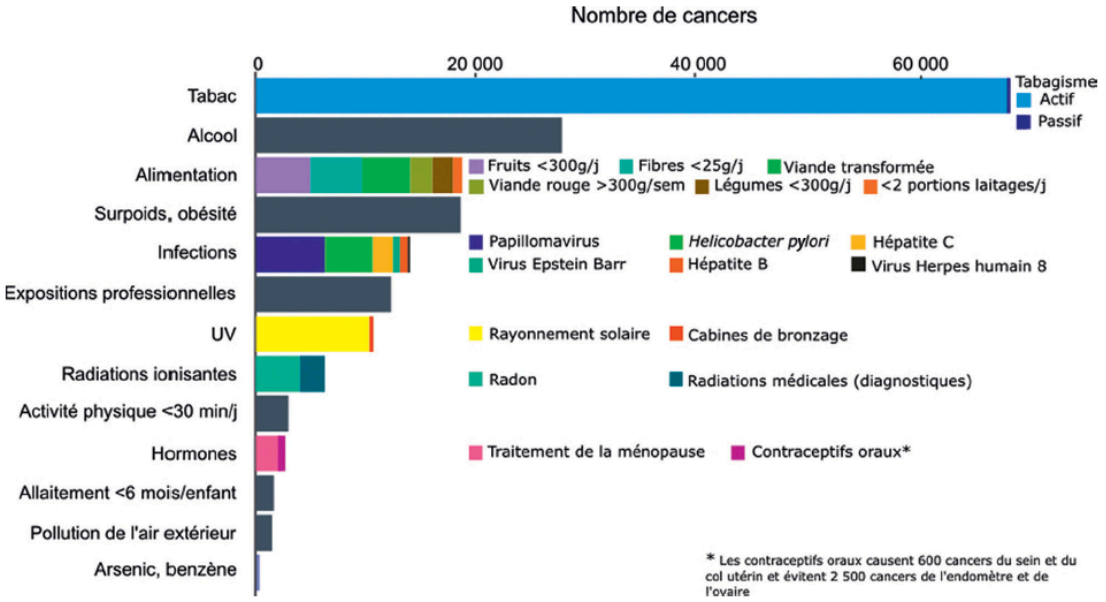


Figure II.1. **Number of new cancer cases attributable to lifestyle and environmental factors among adults aged 30 and over in France, 2015**
Adopted from IARC[139].

2.1 THE "BAD LUCK" DEBATE: STEM CELL DIVISIONS, DRIVER MUTATIONS AND CANCER RISK

Since 2015, Tomasetti and Vogelstein have published a number of papers[9,140–143] in which they studied factors influencing the development of cancer and, in particular, the role of unavoidable stochastic factors that were then popularized as "bad luck". Their starting point is the strong correlation ($R^2 \cong 2/3$) observed between the lifetime cancer risk for different types of tissues and the total number of lifetime stem-cell divisions (LSCD) in such tissues as estimated by a mathematical method they developed. They advanced the thesis that the cause for this correlation are the driver gene mutations that randomly occur during these divisions and that represent the necessary events leading to cancer. By observing that on average tissues with a higher number of lifetime stem-cell divisions present a higher cancer risk they suggested that an intrinsic and unavoidable stochastic risk factor has a major role in cancer development.

As LSCDs are not relevant for this thesis, the mathematical model developed by Tomasetti and Vogelstein for estimating the total number of LSCD in a tissue and its limitations will not be discussed in detail. Here, we simply recall that this model depends on two parameters: the number *s* of stem cells found in fully developed tissues and the total number *d* of divisions each of these cells undergo in the lifetime of an individual. After estimating LSCD for 25 different tissues for which parameter estimates are available, the two authors showed that the observed correlation between lifetime cancer risk (CR) in the US and the LSCD is 0.81 which implies that the proportion of the variation of log(CR) explained by log(LSCD) is R^2=0.66 [=0.81^2]. They found similar correlations using CR figures from 68 different countries.

Unfortunately, this result was misrepresented as if "2/3 of new cancer cases" were due to "bad luck". This provocative interpretation is wrong because 2/3 refers to cancer risk in tissue types and therefore it says nothing about the probability of an individual to develop cancer. Moreover, it is not possible to interpret this correlation as a measure of the fraction of risk attributable to some risk factor[144]. These results and their misinterpretation by some of the media sparked a debate about the role of randomness in cancer; several authors expressed serious concerns about the potential danger that inaccurate interpretation and dissemination of such statistical findings could bring to primary prevention[140].

In a subsequent paper published in 2017, the two authors provided a clearer conceptual distinction between the proportion of preventable cancers and the proportion of driver mutations due to environmental factors and, using cancer genome sequences and epidemiological data, estimated the proportions of driver gene mutations due to environmental (E), hereditary (H) and replicative factors (R), the latter being intrinsic random factors. In particular, they estimated the number of mutations due to R from genomic sequences from "unexposed" individuals, while genomic sequences from exposed individuals were used to estimate the total number of mutations. Even though in principle partitioning

causes in this way is inaccurate and unrealistic as R is likely to be modulated by the environment or the genetic background, this approach has the advantage of establishing a quantifiable link between the proportion of preventable cancers and the proportion of driver mutations due to E through a model relating them to the relative risk and the prevalence of the environmental factor E. To understand this, Tomasetti and Vogelstein proposed the conceptual example illustrated in Figure II.2, where three driver mutations are the necessary condition to develop cancer. Consider a cohort of 20 individuals with cancer, where all individuals have the three mutations and all but two are exposed to a carcinogenic exposure, such as cigarette smoking.

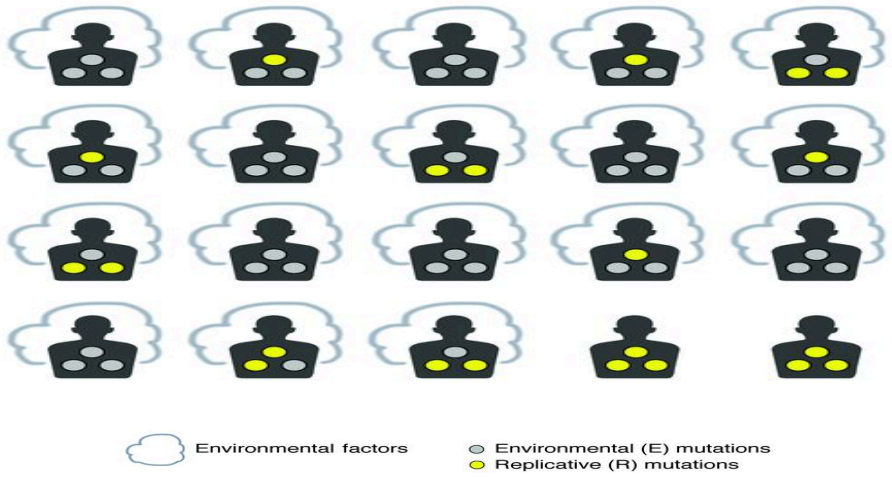


Figure II.2. **Mutation aetiology in lung adenocarcinoma**
Modified from Tomasetti and colleagues[143].

In the example depicted in the figure, driver mutations due to the environment E are in grey and those due to intrinsic random factors (replications, R) are in yellow, so that E accounts for 21/60=35% of the driver mutations in the population and R for 39/60=65% of them. Even though intrinsic random factors have thus a predominant role, 18/20=90% of new cases could be prevented by eliminating environmental factors: if we removed E, all grey mutations would disappear and only two individuals would remain with all the three mutations, all due to intrinsic random factors, that would lead to cancer.

This illustration shows that chance might have a large role in the appearance of deleterious mutations and yet the majority of cases could be prevented by eliminating exposure. As a matter of fact, even if cancer is known to be caused by uncontrolled cell divisions, the main biological cause of the disease remains poorly understood. As argued by Kelly-Irving and colleagues[145], random occurrences of mutations do not equate to random occurrences of cancer and mutation is a necessary, but not sufficient condition for the development of cancer.

To put this debate into context, we note that in addition to the somatic mutation theory previously discussed in chapter I (accumulation of somatic mutations in oncogenes and tumour suppressor genes leading to cancer development), a stem cell division theory of cancer (SCDTC) has been advanced more

recently. According to such theory, the risk of developing cancer is not only increased by mutagenic factors, but also by any factor that promotes the accumulation of cell divisions in stem cells by acting on the stem cell or on the stem cell environment such as physiological changes in the levels of hormones and growth factors, cell death occurring during physiological tissue renewal, cell death (or cellular damage) occurring during pathological conditions (e.g. tissue injury, inflammation and infection), and exposure to non-mutagenic environmental factors [26] (Figure II.3).

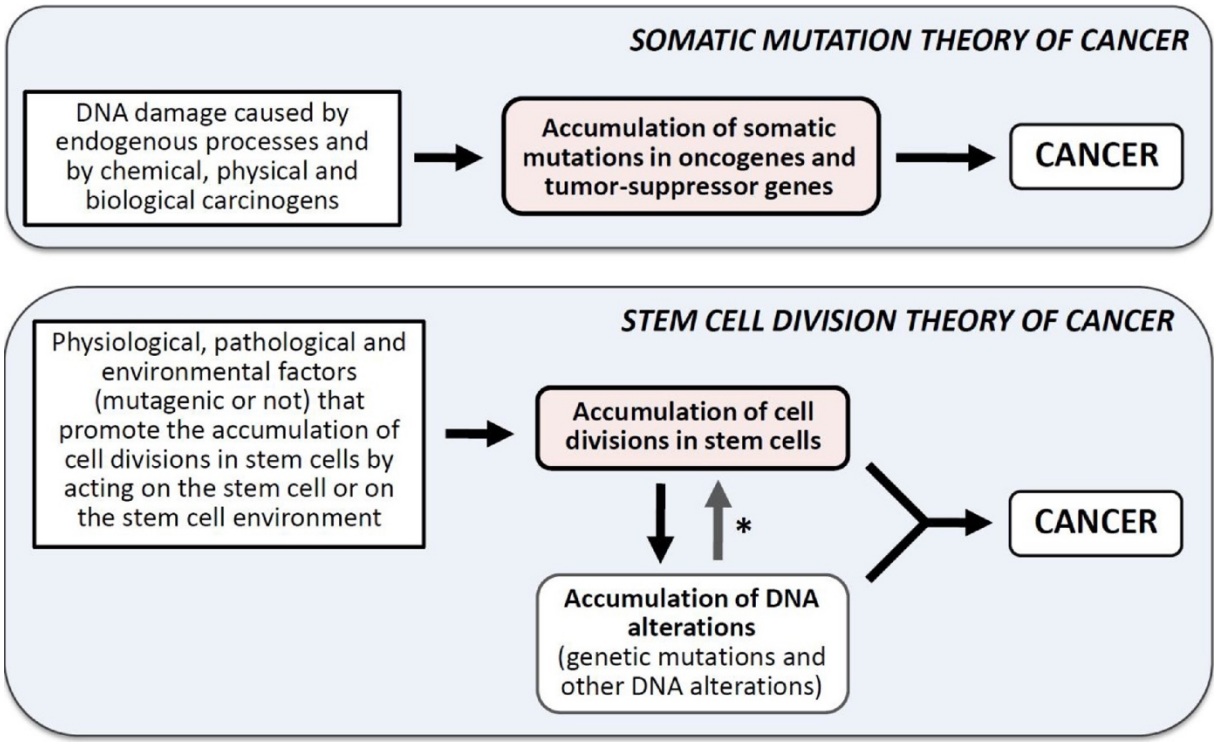


Figure II.3. **Somatic mutation and stem cell division theories of cancer**
Adopted from Lòpez-Làzaro[146]

2.2 PREDICTING LUNG CANCER RISK VIA *EXTRINSINC* MUTATIONS

Wu and colleagues proposed an alternative method to estimate the proportions of mutations due to intrinsic and extrinsic factors that is based on mutational signatures [147]. As COSMIC signature 1 correlates with age at cancer diagnosis, Wu and colleagues used the ratio between the number of mutations associated with such signature and the total mutation burden as a proxy for the proportion of intrinsic mutations. Using this approach, they estimated that the vast majority of mutations (70%-90%) is due to extrinsic factors in most cancer types, a result that contradicts the findings of the 2017 paper by Tomasetti and Vogelstein.

We adopted a similar approach based on the use of mutational signatures to address the issue of "bad luck" and preventable cancers. Collaborators used genome sequences data and extracted mutational signatures obtained from previous research[118], to estimate mutation rates caused by tobacco smoking in different tissue types.

We then compared such estimated mutation rates to cancer incidence hazard ratios and mortality rates in smokers and non-smokers in the same tissues. As shown in Table II, the correlation between mutation rates in smokers and cancer incidence hazard ratios for smokers relative to non-smokers is much more evident than the association of the latter with the stem cell lifetime divisions estimated by Tomasetti and Vogelstein.

In particular, the correlation between the cancer incidence hazard ratio for smokers relative to non-smokers and the mutation rates (per pack-year) in smokers is strong (ρ=0.93, p=2× 10^{-2}). The correlation becomes negative and weaker (ρ=-0.65, p=2.3× 10^{-1}) when we compare the cancer incidence hazard ratio for smokers with the cumulative stem cell divisions (Table II).

The pattern for former smokers is similar, with a strong correlation between the cancer incidence hazard ratios and mutation rates per pack-year (ρ=0.91, p=3× 10^{-2}), while cumulative stem cell divisions are only weakly negatively correlated with cancer hazard ratios (ρ=-0.58, p=3.1× 10^{-1}). Similar findings are obtained when mortality rates are used instead of cancer incidence rates, although none of the correlation coefficients were significantly different from zero (all p>1× 10^{-1}).

Our results reinforce the findings from Little and colleagues[148] that using data taken from the 2015 Science paper of Tomasetti and Vogelstein concluded that stem cell divisions are poorly predictive of smoking-related risk.

Table II. **Comparison between mutation rates, cumulative stem cell lifetime divisions, hazard ratios (HR) for cancer in smokers and mortality rates in smokers and never smokers, for the cancer sites for which information was available in all sources**
Adopted from Perduca and colleagues[24]

Cancer site	Mutation rates in smokers [a]	Cumulative stem cell lifetime divisions [b]	Incidence HR for smoking men [c]	Incidence HR for former smoking men [c]	Mortality rates smokers with ≥25 cigarettes/day /non-smokers [d]
Lung adenocarcinoma	150.5	9.272×10^{9} [e]	23.30	5.28	415.2 / 16.9
Larynx	137.7	3.186×10^{10} [f]	13.24	3.51	17.3 / 0
Pharynx	38.5	NA	6.67	2.06	19.4 / 0
Bladder	18.3	NA	3.84	2.15	51.4 / 13.7
Esophagus (squamous)	N.S.	1.203×10^{9}	3.94	1.26	50.0 / 5.7
Liver	6.4	2.709×10^{11}	2.92	2.09	31.3 / 4.4
Pancreas adenocarcinoma	N.S.	3.428×10^{11}	1.62	0.89	52.9 / 20.6

[a] Statistically significant average number of somatic substitutions per genome per pack-year[118]
[b] Cumulative number of divisions of stem cells per lifetime. From Tomasetti and Vogelstein[9]
[c] HRs relative to non-smokers. From Agudo and colleagues[149]
[d] Cumulative mortality rate per 100,000 persons per year[150]
[e] Cumulative number of divisions of stem cells per lifetime[9]
[f] Adenocarcinoma (same rate in smokers and non-smokers)

3. CONCLUSION

Understanding how cancer develops is crucial for improving prevention strategies. It is well accepted that carcinogens leave fingerprints (traces of past events, including the action of environmental factors). The mutational and epigenetic profile of a cancer genome result respectively from the superposition of all the traces, or signatures, left by mutational processes and the alteration of methylation levels due to environmental, lifestyle (and random) factors. Both types of signatures represent promising areas of research that are likely to continue to contribute novel insights into the nature of cancer and the processes that lead to it. Such gains in new knowledge are likely to accelerate when epidemiological studies are going to routinely collect and sequence DNA from tumour tissue allowing the analysis of mutational signatures and the linking of such signatures to epidemiological data.

According to the prevailing model of carcinogenesis, cancer is primarily caused by the accumulation of genetic mutations. However, it is increasingly accepted that the accumulation of somatic mutations alone cannot explain the development of cancer. Evidence is accumulating that genetic and non-genetic mechanisms such as epigenetic alterations and environmental factors may influence stem-cell divisions and therefore cancer development. In this respect, it would be very interesting to try to estimate the effect of such factors on the number of lifetime stem cell divisions. This would require building a model for estimating the fraction of such events over the total number of events required for cancer development. Other events or conditions that may play an important role but have not yet been considered in the model of cancer development are disrupted or inefficient DNA repair mechanisms, that may be limited to some organs, and dysfunctions of immune surveillance.

CHAPTER III:

COMPUTATIONAL TOOLS TO DETECT SIGNATURES OF MUTATIONAL PROCESS

This chapter will cover one of the most recent developments with regards to cancer genomics: the identification of mutational signatures from cancer genomes that may be linked to specific exogenous and endogenous factors responsible for the development of cancer. This field is growing rapidly and is leading to strong collaborations between quite diverse disciplines and in particular genomics, bioinformatics, biostatistics and epidemiology. Major international projects such as Mutograph funded by CRUK are collecting at the same time extensive epidemiological data as well as tumour DNA that is then sequenced in order to try to link mutational signatures to specific exposures. In this work we focused on the large number of analytical methods and tools that have been developed in the last few years to extract and identify mutational signatures from sequencing data from tumour DNA. We introduce a probabilistic model for simulating mutational catalogues and we exploit it to produce an original empirical comparison of the performance of most of the currently available tools for the analysis of mutational signatures.

Contribution
First author, discussed the analytical strategy with the supervisors, conducted statistical analyses, wrote the first draft of the manuscript and replied to reviewers' comments.

1. Context **87**
2. Overview of avalaible tools for mutational signature analysis **88**
2.1 *De novo* approaches **92**
2.2 Refitting with known mutational signatures **93**
2.3 Combining de novo and refitting procedure **94**
3. Materials and experimental settings **95**
3.1 The Cancer Genome Atlas **95**
3.2 Our original refitting tool: MutationalCone **95**
3.3 Simulation of a mutational catalogue **96**
4. Comparison of algorithms performance **100**
4.1. Specificity and sensitivity for *de novo* extraction and assignment **100**
4.2 Bias of refitting procedures **101**
5. Findings **102**
5.1 Performance of *de novo* tools **102**
5.2 Performance of refitting tools **110**
6. Conclusion **113**

1. CONTEXT

After the introduction of the original framework for the identification of mutational signatures, several other mathematical methods and computational tools have been proposed for their detection and for the estimation of their contribution to a given catalogue. As reported in chapter I, these methods can be grouped in two categories with different goals. The first class of methods aims to discover novel signatures while the second class aims to detect the known and validated mutational signatures in the mutational catalogue of a given sample. The approaches used in the first class are referred to as "*de novo*" (or "signature extraction") while those in the second class as "refitting" (or "signature fitting"). All methods have been implemented in open source tools, mainly R packages, but some of them are available through command line, the Galaxy project or a web interface.

Signatures identified with de novo methods can be compared to reference signatures (for instance those listed in COSMIC) through measures such as cosine[16] or bootstrapped cosine similarity[15], which is a distance metric between two non-zero vectors. In this step of the analysis, extracted signatures are matched to the most similar reference signature, provided that their similarity is greater than a fixed threshold.

To date, more than twenty methods with similar aim (minimize the distance between original mutational catalogue and the estimated one) are available. However, no systematic evaluation of the performance of these methods has been conducted and the issue of the choice of an appropriate cosine similarity threshold when matching a newly extracted signature to the most similar counterpart in a reference set has not been addressed yet.

2. OVERVIEW OF AVALAIBLE TOOLS FOR MUTATIONAL SIGNATURE ANALYSIS

A similar number of de novo and refitting methods exist and all of them are available as open source tools, mainly as R packages, or web interfaces (Table III). The typical input of these tools is a file including the mutation counts but some tools derive the mutation counts from ad-hoc input files that may include for each individual a list of mutated bases, their position within the genome and the corresponding bases from a reference genome. The typical format of such input files is MAF, Variant Call Format (VCF) or less common formats such as (Mutation Position Format) MPF and Mutation Feature Vector Format (MFVF).

For biologists or those who are not familiar with programming, a set of tools were also developed and provided with user-friendly interfaces. Some tools include additional features such as the possibility to search for specific patterns of mutations (e.g. APOBEC-related mutations[16]) and differential analysis[151].

Table III. **Available tools for the detection of mutational signatures.**

Software	Available platform/model	Input files	Additional features
		***de novo* approaches**	
WTSI[11]	MATLAB/ NMF		- Original framework - An improved version has been recently implemented in SigProfiler
EMu[13] https://github.com/andrej-fischer/EMu	Command line/EM algorithm	- Mutation counts file - With respect to other tools, the counts file is transposed (the rows correspond to the samples)	- Opportunity matrix - Selection of the optimal number of signatures
SomaticSignatures[152] https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html	R/NMF and PCA	Variant Call Format	- Group-wise comparisons - Genomic visualization - Hierarchical clustering
pmsignature[153] https://github.com/friend1ws/pmsignature	R/mixed-membership model	- Mutation Position Format - Mutation Feature Vector Format	- Reduction of complexity - Mutation types defined by one or two flanking bases - Selection of the optimal number of signatures - Transcriptional strand bias - Background signature
bayesNMF[21,154–156] https://github.com/jburos/bayesNMF https://software.broadinstitute.org/cancer/cga/msp	R/Bayesian NMF	Mutation counts file	- Selection of the optimal number of signatures - Data pre-treatment with the function get.lego96.hyper reduces the influence of hypermutated catalogues
signeR[151] https://bioconductor.org/packages/release/bioc/html/signeR.html	R/Bayesian NMF	Variant Call Format	- Opportunity matrix - Selection of the optimal number of signatures - Group-wise comparison (differential analysis)
mutSignatures[157] https://cran.r-project.org/web/packages/mutSignatures/index.html	R/NMF	Mutation counts file	- R-based implementation of WTSI[11]
maftools[16] https://bioconductor.org/packages/release/bioc/html/maftools.html	R-Bioconductor /NMF	- Mutation Annotation - Format	- Genomic visualization - Cosine similarity - Selection of the optimal number of signatures - Group-wise comparisons (differential analysis) - APOBEC enrichment analysis

Continued on the following page

Helmsman[158] https://github.com/carjed/helmsman	Python/ NMF and PCA	- Variant Call Format - Mutation Annotation Format	- Able to run in parallel and designed for large datasets - Connection to external packages (in R) - may generate mutational catalogues from sequence data
SignatureAnalyzer[21] https://www.synapse.org/#!Synapse:syn11801492	R/ Bayesian NMF	Mutation counts file	- Automatic selection of the optimal number of signatures - Sparse signature profiles and contributions
SigProfiler[11,21] https://fr.mathworks.com/matlabcentral/fileexchange/38724-sigprofiler	Matlab/ NMF	Mutation counts file	- Further development of the original framework - Two steps: 1) extraction of a minimal set of signatures, 2) estimation of their contributions to individual samples
SparseSignatues[159] https://bioconductor.org/packages/release/bioc/html/SparseSignatures.html	R/ NMF with Lasso-penalized cost function	Mutation counts file	- Integration of DNA replication error signature - Sparse signature matrix - Number of signatures estimated with cross-validations - Scalable to large datasets
		Refitting approaches	
deconstructSigs[160] https://github.com/raerose01/deconstructSigs	R/linear regression	Mutation counts file	- Opportunity matrix
Qpsig[161] https://f1000researchdata.s3.amazonaws.com/supplementary/8918/0d25c07c-16ba-4b14-91e7-71749dcbbdd5.pdf	R/quadratic programming	Mutation counts file	
SignatureEstimation[162] https://www.ncbi.nlm.nih.gov/CBBresearch/Przytycka/index.cgi\#signatureestimation	R/quadratic programming and simulated alienation	Mutation counts file	
MutationalPatterns[163] http://bioconductor.org/packages/release/bioc/html/MutationalPatterns.html	R/Non-Negative Least Squares	Mutation counts file	- Also de novo identification - Cosine similarity comparison - Strand bias analyses - Enrichment and depletion
YAPSA[164] http://bioconductor.org/packages/release/bioc/html/YAPSA.html	R/Linear Combination Decomposition	Mutation counts file	- Cut-off for normalized exposure - Enrichment and depletion
decompTumor2Sig[165] https://github.com/rmpiro/decompTumor2Sig	R/quadratic programming	- Variant Call Format - Mutation Position Format - Mutation Feature Vector Format	- Converts a set of "Alexandrov's signatures"[8] to "Shiraishi's signatures"[153] - Decomposes a mutational catalogue in "Shiraishi's signatures"

Continued on the following page

MutationalCone [Appendix 2]	R/cone projection	Mutation counts file	- Fast in comparison to others refitting tools
Sigfit[166] https://github.com/kgori/sigfit	R/ Bayesian NMF	Mutation counts file	- Provides a new model for combining de novo and refitting approaches - Possible application to indel or rearrangement count data - Also implements EMu[13] model and allows conversion to genome-or exome- relative signatures
	Pipelines and web-interfaces		
Mutspec[167] https://toolshed.g2.bx.psu.edu/repository/view_repository?id=f5c1f75e9fb33f8e	Galaxy pipeline/NMF	Variant Call Format	- de novo identification - Includes MS analysis in mouse cancer
MutaGene[168] https://www.ncbi.nlm.nih.gov/research/mutagene/	Web-interface	TCGA and ICGC data	- Refitting and de novo identification - Clustering of samples according to mutational profiles - Identification of potential driver's mutations
mSignatureDB[15] http://tardis.cgu.edu.tw/msignaturedb/	Web-interface	- Variant Call Format - Mutation Annotation Format - TSV	- Refitting and de novo identification - Bootstrapped cosine similarity - Comparison with either hg19 or hg38
Mutalisk[169] http://mutalisk.org	Web-interface	Variant Call Format	- Refitting and de novo identification - Transcriptional strand bias - Localization of kaetegis - Histones modifications - Cosine similarity comparison
MuSiCa[170] http://bioinfo.ciberehd.org:3838/MuSiCa/	Web-interface	- Variant Call Format - Mutation Annotation Format - TSV - Excel	- Refitting and de novo identification - Cosine similarity - Samples classification

2.1 *DE NOVO* APPROACHES

Most tools that have been developed to identify mutational signatures were based on decomposition algorithms including NMF or a Bayesian version of NMF. The original method developed by Alexandrov et al. was based on NMF and was implemented in MATLAB[11] and is available also as an R package developed independently[157]. An updated and elaborated version named SigProfiler, was proposed recently for extracting a minimal set of signatures and estimating their contribution to individual samples[21]. The latter article also discusses an alternative method based on Bayesian NMF, called SignatureAnalyzer, that led to the identification of 49 reference signatures. Another tool that utilizes NMF is maftools that is one of the few *de novo* tools that allows systematic comparison with the 30 validated signatures in COSMIC by computing cosine similarity and assigning the identified signatures to the COSMIC one with the highest cosine similarity[16].

Other tools such as SomaticSignatures[152] or the recent Helmsman[158] allows the identification of mutational signatures through Principal Component Analysis (PCA) in addition to NMF. For the sake of our formal comparison of the tools' performance, we have only tested NMF implementations because in PCA the factors are orthogonal and the values inside the matrix can potentially be null or negatives, which is a deviation from the paradigm postulating that catalogues are the superposition of positively weighted signatures. However, PCA could be a promising way to explore complex situations in which mutational processes interfere with each other (e.g. relatively error free repair processes competing with error prone repair processes). Developed in the Python language, Helmsman allows the rapid and efficient analysis of mutational signatures directly from large sequencing datasets with thousands of samples and millions of variants.

SparseSignatures[159] proposes an improvement of the traditional NMF algorithm based on two innovations, namely the default incorporation of a background signature due to DNA replication errors and the enforcement of sparsity in identified signatures through a Lasso penalty. This latter feature allows the identification of signatures with well-differentiated profiles, thus reducing the risk of overfitting.

In addition to decomposition methods, an approach based on the Expectation Maximization (EM) algorithm has been proposed to infer the number of mutational processes operative in a mutational catalogue and their individual signatures. This approach is implemented in the EMu tool[13], where the underlying probabilistic model assumes that input samples are independent and the number of mutational signatures is estimated using the Bayesian Information Criterion (BIC). Another tool that uses a probabilistic model named mixed-membership model is pmsignature[153]. This tool utilizes a

flexible approach that at the same time reduces the number of estimated parameters and allows to modify key contextual parameters such as the number of flanking bases.

The latter feature may be particularly useful as the standard and most commonly used methods based on trinucleotides may not be the most adequate to detect specific mutational processes that lead to larger-scale substitution patterns. Evaluating the impact of limiting to trinucleotides or estimating the gain in performance associated with the extension of the context sequence to two flanking bases, is difficult and beyond the scope of our work. However, it is worth noting that trinucleotide-based methods have been able to identify several signatures associated with defective DNA mismatch repair and microsatellite instability (i.e. signatures 6, 14, 15, 20, 21, 26 and 44 of COSMIC v3)[21]. It is important to note that for the purpose of the comparison with the other tools, the number of flanking bases was set to one, and therefore we considered 96 mutation types.

EMu, signeR and pmsignature (and the refitting tool deconstructSigs) have been designed to take into account the distribution of triplets in a reference exome or genome, for example from a sequence of normal tissue in the same individual. This is done by "normalizing" the input mutational catalogues with respect to the distribution of triplets in the reference exome or genome using an "opportunity matrix".

2.2 REFITTING WITH KNOWN MUTATIONAL SIGNATURES

In addition to the identification of novel mutational signatures, scientists are often interested in evaluating whether a signature observed in an individual tumor belongs to an established set of signatures (e.g. the COSMIC signatures). This task is performed by "refitting tools" that aim to search for the "best" combination of established signatures that explains the observed mutational catalogue by projecting the latter into the multidimensional space of all non-negative linear combinations of the N established signatures.

The deconstructSigs[160] tool searches for the best linear combination of the established signatures through an iterative process based on multiple linear regression aimed at minimizing the distance between the linear combination of the signatures and the mutational catalogue. All the other tools minimize the distance through equivalent approaches based on quadratic programming[161,162,165], non-negative least square[163] linear combination decomposition [164] and simulated annealing[162].

2.3 COMBINING DE NOVO AND REFITTING PROCEDURE

Sigfit[166] is a recently introduced R package for Bayesian inference based on two alternative probabilistic models. The first of such models is a statistical formulation of classic NMF where signatures are the parameters of independent multinomial distributions and catalogues are sampled according to a mixture of such distributions with weights given by the exposures, while the second model is a Bayesian version of the EMu model. An interesting innovation of Sigfit is that it allows the fitting of given signatures and the extraction of undefined signatures in the same Bayesian process. As argued by the authors, this unique feature might be helpful in cases where the small sample of catalogues makes it difficult to try to identify new signatures or when the aim is to study the heterogeneity between the primary tumor and metastasis in terms of the signatures they show.

In this work, we empirically evaluate the methods that have been already presented in a peer review published paper to date and for which an implementation in R is available. To this aim, we adopt the COSMIC set as reference for the analysis of simulated and real mutational catalogues because we evaluate tools that were developed at the time when COSMIC was the only available database of reference.

3. MATERIALS AND EXPERIMENTAL SETTINGS

3.1 THE CANCER GENOME ATLAS

In order to evaluate the performance of the available algorithms on real data, exome sequences from The Cancer Genome Atlas (TCGA) repository (https://cancergenome.nih.gov/) were used for four cancer types: breast cancer, lung adenocarcinoma, B-cell lymphoma, and melanoma.

Mutation Annotation Format (MAF) files with the whole-exome somatic mutation datasets from these cohorts were downloaded from the portal gdc.cancer.gov on 6 March 2018. Data were annotated with MuSE[171] and the latest human reference genome (GRCh38). Mutational catalogues from these cohorts were obtained by counting the number of different mutation types using maftools[16]. The distribution of the number of mutations for each sample and separately for each cancer type is depicted in Figure III.1.

According to the COSMIC website, 13 and 7 signatures have been found for breast cancer and lung adenocarcinoma respectively, 6 for B-cell lymphomas and 5 for melanoma.

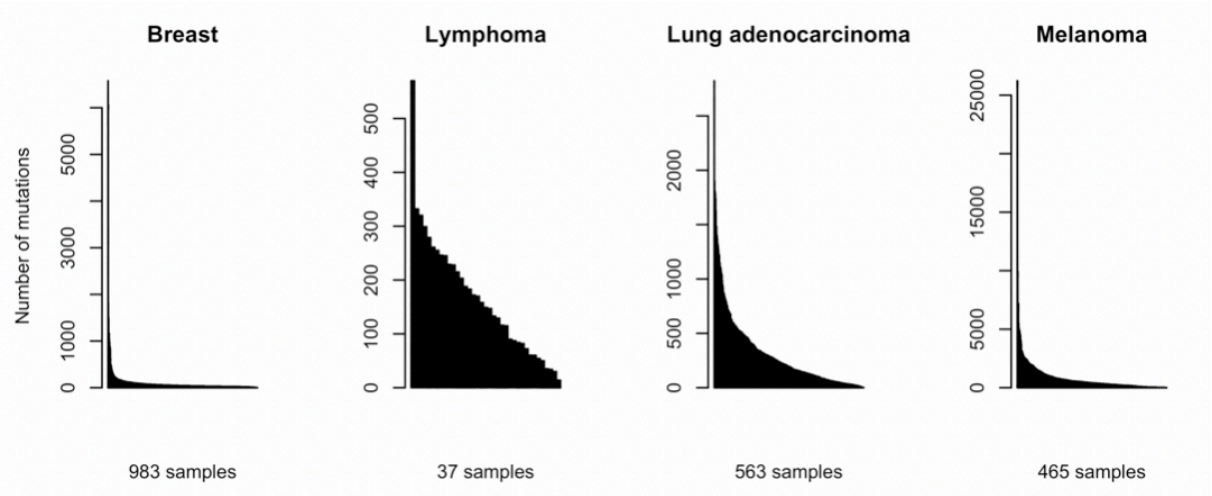


Figure III.1. **Barplot with the number of mutations in each sample in four TCGA cohorts. Each bar represents a sample, with the number of mutations shown in the y-axis.**

3.2 OUR ORIGINAL REFITTING TOOL: MUTATIONALCONE

We propose an alternative implementation of the decomposition performed by Huang[162] or Huebschmann[164] based on a simple geometric framework. Finding the linear decomposition of the input catalogue M on a set of given signatures minimizing the distance can be seen as the problem of projecting M on the geometric cone whose edges are the reference signatures. We propose to solve this problem by applying the very efficient R package called coneproj [172]. More details about our algorithm, which we called MutationalCone, together with the R code implementing it, can be found in Appendix 2.

3.3 SIMULATION OF A MUTATIONAL CATALOGUE

The first key assumption of our original model for the simulation of mutational catalogues is that the number of mutations in a sample g that are induced by process n follows a zero-inflated Poisson (ZIP) distribution. According to this two-component mixture model, e_g^n is either 0 with probability π or is sampled according to a Poisson distribution $\mathcal{P}(\lambda)$ with total probability $1-\pi$. Such a model depends on two parameters: the expectation λ of the Poisson component, and the probability π of extra "structural" zeros. The ZIP model allows for frequent zeroes and is therefore more suitable for modelling a heterogeneous situation where some samples are not exposed to a given mutational process ($e_g^n = 0$) while some others are ($e_g^n > 0$). Realistically, the mutation counts due to process n in each of the G samples, $e_1^n, \dots, e_G^n$, are assumed to be independent and identically distributed according to a ZIP model where the expectation of the Poisson component is specific to n:

$$e_g^n \sim ZIP(\lambda_n, \pi), \text{ for all } n = 1, \dots, N.$$

Note that the expected number of mutations in sample g due to process n is $(1-\pi)\lambda_n$. This flexibility given by process-specific average counts is the second important characteristic of the model and reflects the possibility that the mutagenic actions of different processes are intrinsically different with respect to their intensity. Obviously, it would have been possible to do one step further and allow for parameters $\lambda_{n,g}$ specific to both processes and samples, thus representing the realistic situation in which the exposures of different samples to the same process have different duration or intensity (e.g. smokers/non-smokers). However, this would have resulted in too many parameters to tune, thus making it difficult to interpret the results of our simulation study. For the same reason we considered one fixed value of π.

The parameter λ_n depends on both the average total number r of mutations in a sample and the relative contribution of n. We therefore imposed the parameterization $\lambda_n(1-\pi) = q_n r$, where q_n is the average proportion of mutations due to the process n.

When taking a unique value of r, this model produces realistic simulations even though it underrepresents extreme catalogues with very large or small total numbers of mutations (Figure III.2 (c)). While considering a specific value of r for each sample, or group of samples, would definitely make it possible to obtain a more realistic distribution of simulated catalogues Figure III.2 (b)) such multidimensional parameter would complicate unnecessarily the empirical assessment of mutational signature detection methods by introducing too many specifications. Therefore, a unique r was considered for each set of simulations. This formulation allows to study empirically the performance of a given signature detection method as a function of the average number of mutations r while fixing the average proportion of mutations due to each mutational process q_n, according to different profiles that

mimic real cancer catalogues. Interestingly, the ZIP model appeared to be more appropriate to represent mutational catalogues than the pure Poisson model used in previous publications [13,151] (Figure III.2 (d))).

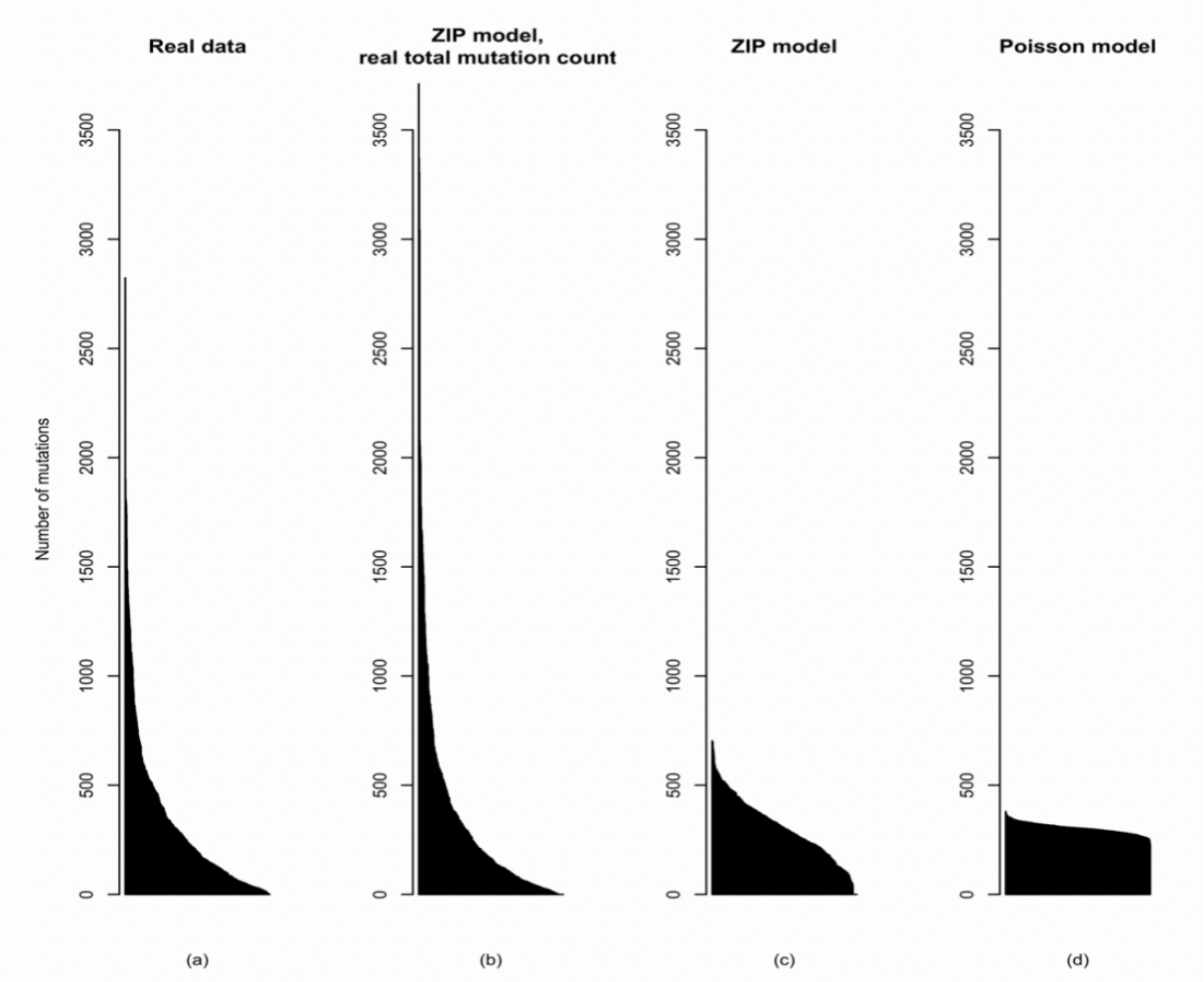


Figure III.2. **Simulations of 563 lung adenocarcinoma catalogues according to different models**

(a) Real catalogues from the TCGA lung adenocarcinoma cohort. (b)-(c) Catalogues sampled from the ZIP model described in the main text. The relative contribution q_n of each signature n is the mean of the relative contributions of n in all samples as estimated by maftools. In (b) simulated and real catalogues are in a 1 to 1 correspondence: for each simulated sample g, the total number of mutations r_g in the corresponding real catalogue is taken. In (c) all samples are simulated according to $r = 306$, the average total number of mutations in the real data. The latter example illustrates the parametric model used for the simulation study. (d) Catalogues sampled according to the Poisson model $e_g^n \sim P(\lambda_n)$, where λ_n is the mean number of mutations due do to n in the real samples as estimated by maftools.

We adopted the following simulation protocol:

1. We chose N signatures from the COSMIC database, thus obtaining the matrix P.

2. For each sample g and process n, we sampled e_g^n from a ZIP distribution with parameters $\lambda_n = q_n r/(1-\pi)$ and π and obtained E. Here q_n, r and π are fixed parameters to set.

3. Then, we computed the product $P \times E$. In order to obtain the final simulated catalogue M, some noise was added to the latter matrix by taking $m_g^k \sim \mathcal{P}\left((P \times E)_g^k\right)$.

Four alternative sets of simulated catalogues were generated, referred to as Profiles 1, 2, 3 and 4, each set mimicking a particular cancer: breast cancer, lung adenocarcinoma, B-cell lymphoma and melanoma. In order to do so, for each tumour type, we applied MutationalCone to the corresponding TCGA datasets and we calculated the mean contribution across all samples of each signature known to contribute to the specific cancer type q_n. Signatures with $q_n = 0$ do not contibute to the final catalogue and were not in the matrix P. Figure III.3 depicts the resulting four sets of configurations $(q_1, \dots, q_N)$ used for the simulations. Profiles 3 and 4 are characterized by one dominant signature, Profiles 2 by two signatures with similar large contributions and Profile 1 by several signatures with small effects.

Four different configurations $(q_1, \dots, q_{30})$ were considered for simulating realistic data. Each configuration represents the average share of mutations due to the different COSMIC signatures and was chosen to mimic real exposure profiles for four cancer types: estimates were obtained from Breast Cancer (Profile 1), Lymphoma (Profile 2), Lung Adenocarcinoma (Profile 3) and Melanoma (Profile 4) TCGA cohorts.

The relative frequency of structural zero contributions to the catalogues was fixed to $\pi = 0.6$ in all simulations. This value was chosen because it leads to a small number of hypermutated catalogues, as it is often encountered in practice. Finally, the number of r was set from as little as 10 to as much as 100,000 mutations. This allowed us to study the performance of methods on a large spectrum of catalogues: from a limited number of mutations as in exomes, to a very large number, as in whole cancer genome sequences.

For each of the four tumor types and for each value of r, a catalogue matrix was simulated with G samples.

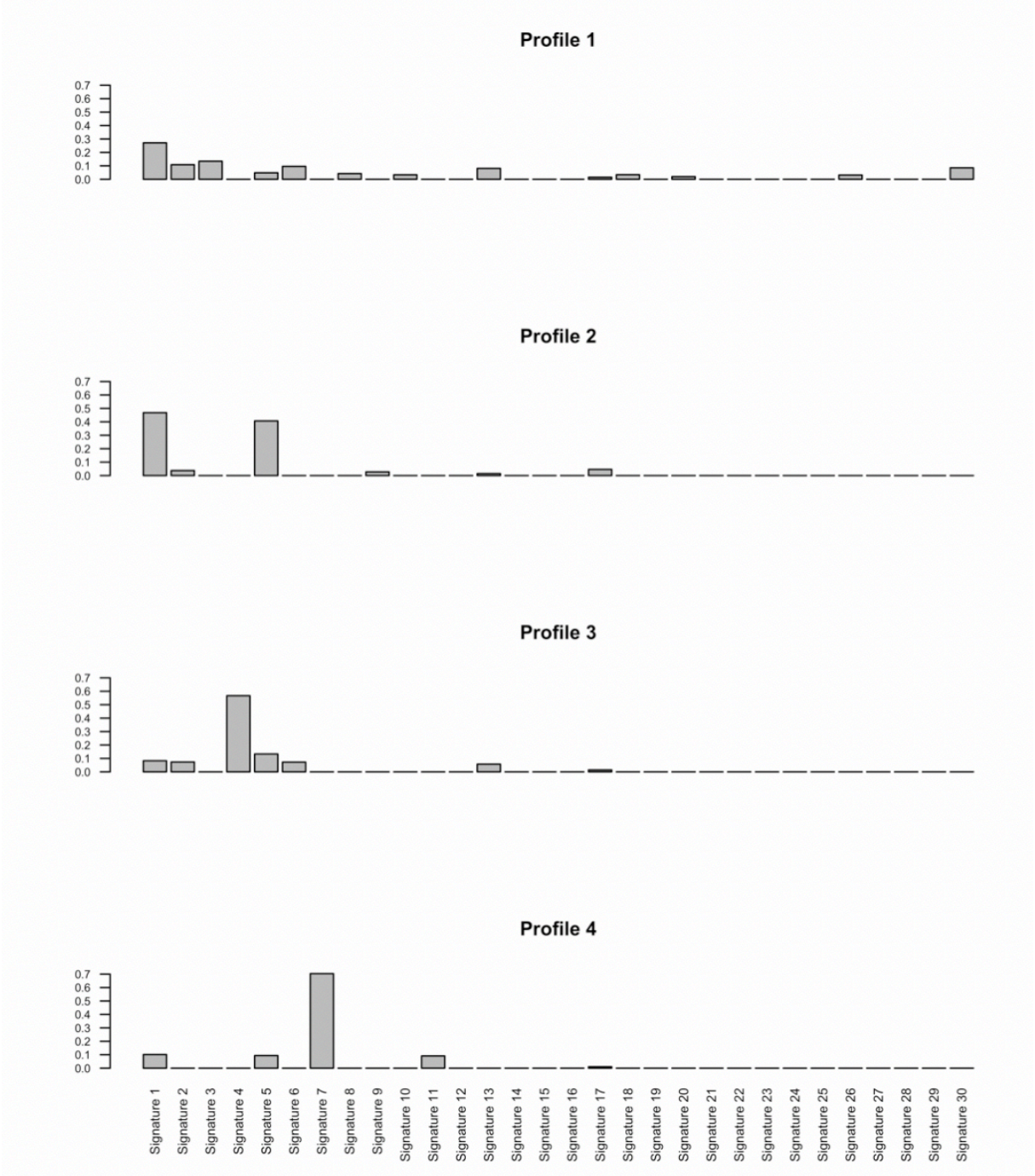


Figure III.3. **Choice of parameters q_n in the simulations.**

Profiles 1-4 respectively mimic real exposure profiles for four cancer types and estimates were obtained from Breast Cancer (Profile 1), Lymphoma (Profile 2), Lung Adenocarcinoma (Profile 3) and Melanoma (Profile 4) TCGA cohorts.

4. COMPARISON OF ALGORITHMS PERFORMANCE

All methods for identifying signatures find solutions to the minimization problem (2). A straightforward way to measure the accuracy of the reconstructed catalogue is, therefore, to calculate the Frobenius norm of the reconstruction error

$$||M - \hat{M}||_F^2 = \sum_{g=1}^{G} \sum_{n=1}^{N} (m_g^k - \hat{m}_g^k)^2,$$

where $\hat{M} = \hat{P} \times \hat{E}$ is the matrix of catalogues reconstructed from the estimated signature and exposure matrices. Some of the algorithms involve stochastic steps such as resampling and/or random draws of initial parameters. For these algorithms, one simple way to assess the robustness of the estimates is to look at the variability of the reconstruction error when the same catalogues are analyzed several times with the same algorithm.

With regards to bayesNMF, it is known that the performance of its principal function might be poor in presence of hypermutated catalogues that mask the detection of signals from less mutated catalogues. For this reason, we pre-treated the catalogues to be analyzed by this tool and replaced hypermutated catalogues by synthetic non-hypermutated catalogues to maintain the original mutational distribution catalogues using the standalone get.lego96.hyper function that can be found in the bayesNMF script.

In order to make decisions about whether an extracted signature is the same as validated signatures (e.g. COSMIC signatures) a cut-off for cosine similarity needs to be defined. We applied six different cut-offs (0,0.75,0.8,0.85,0.9,0.95) and considered as "new" all identified mutational signatures for which the maximal cosine similarity is lower than the cut-off value.

4.1. SPECIFICITY AND SENSITIVITY FOR *DE NOVO* EXTRACTION AND ASSIGNMENT

In most applications, signature extraction is done in two steps: first, signatures are found using a de novo extraction tool and then for the extracted signatures a cosine similarity with each of the COSMIC signatures is calculated. In order to measure the performance of both these steps combined, simulated catalogues were used, and false and true positive rates and false and true negative rates were computed. In a simulated catalogue, the set of *true* signatures $p_1, \dots, p_N$ that do contribute to the catalogue are known, thus allowing the comparison of the latter to the estimated signatures $\hat{p}_{i_1}, \dots, \hat{p}_{i_N}$. Note that, for the sake of simplicity, we set the number of signatures to be found to be equal to the number of signatures used to simulate the catalogues, and thus we do not address questions about model selection performance.

Estimated signatures that belong to the set of "true signatures" are considered as true positives, while all "true signatures" that are not extracted count as false negatives. False positives are all estimated signatures that do not have a match in the set of "true signatures". This can happen for two reasons: the estimated signature is assigned to a COSMIC signature not used to build the catalogue, or it is not sufficiently similar to any COSMIC signature. This last situation usually takes place when setting a very high cosine similarity threshold h. In this case, signatures that have maximal cosine similarity lower than the cutoff, will be termed as "new". Finally, true negatives are all COSMIC signatures not used for the simulation, nor estimated. From these four measures, we compute specificity (number of true negatives divided by the total number of negatives) and sensitivity (number of true positives divided by the total number of positives).

In this empirical study, for each simulation setting described in the Simulated data section (that is for each profile given by a choice of proportions $(q_1, \dots, q_N)$ and for a choice of total number of mutations r) 50 replicates were built, each made of a matrix of G samples. Signatures are then extracted from all replicates with a given tool. Then, extracted signatures are compared to the COSMIC signatures using a cosine similarity threshold h. Finally, we computed specificity and sensitivity and obtained Monte-Carlo estimates based on the means over all replicates.

4.2 BIAS OF REFITTING PROCEDURES

Refitting algorithms assume that the matrix of signatures is known and return the exposure estimates $\hat{e}_g^n$, i.e. estimates of the contribution of each signature e_g^n. A simple way to assess the performance of the refitting method is then to look at the bias of such estimates, by comparing them to the true exposures. In order to do so, we simulated 50 replicates each consisting of one lung adenocarcinoma-mimicking catalogue g (Profile 3) with an average number of mutations set to $r = 10^4$.

Then, for each process n, we obtained Monte-Carlo estimates of the bias $E[\hat{e}_g^n] - e_g^n$ by averaging the differences $\hat{e}_g^n - e_g^n$ over all replicates. A global measure of performance that considers all exposure estimates is given by the mean squared error (MSE), that is the expected value of the loss function $\sum_{n=1}^{30}(\hat{e}_g^n - e_g^n)^2$. We obtained Monte-Carlo estimates of the MSE by averaging the loss function values across all 50 replicates and calculated asymptotic confidence intervals.

5. FINDINGS

5.1 PERFORMANCE OF *DE NOVO* TOOLS

5.1.1 FROBENIUS NORM

Figure III.4 shows the distribution of the reconstruction error when a given computational tool is applied several times to the same real trinucleotide matrix. Reconstruction errors show limited variability due to stochastic steps in the algorithms and no variability whatsoever for maftools. All methods under evaluation are roughly equivalent in terms of their ability to properly reconstruct the initial matrix of mutational catalogues. This is not surprising, given that all methods are meant to solve the optimization problem given in equation (2).

In general, the error value appears to depend on the cancer dataset. This is expected because the fours datasets differ with regards to the number of samples, their total number of mutations and the number of operating mutational signatures, making the decomposition more or less difficult.

Results show that the performance of each method improves after pre-treating the samples, especially for Melanoma and Breast cancer datasets that are characterized by a few samples with an extremely high number of mutations. For the Melanoma dataset, the gain in performance is considerable for bayesNMF and maftools.

Each program under evaluation is applied 50 times on the same matrix of real catalogues shown in Figure III.1; boxplots represent the distribution of the squared Frobenius distance between the original catalogue and its reconstruction. Boxplots look like flat segments because of the scale of the y-axis. Each catalogue was analyzed with or without data pre-treatment with the standalone bayesNMF function get. lego96.hyper.

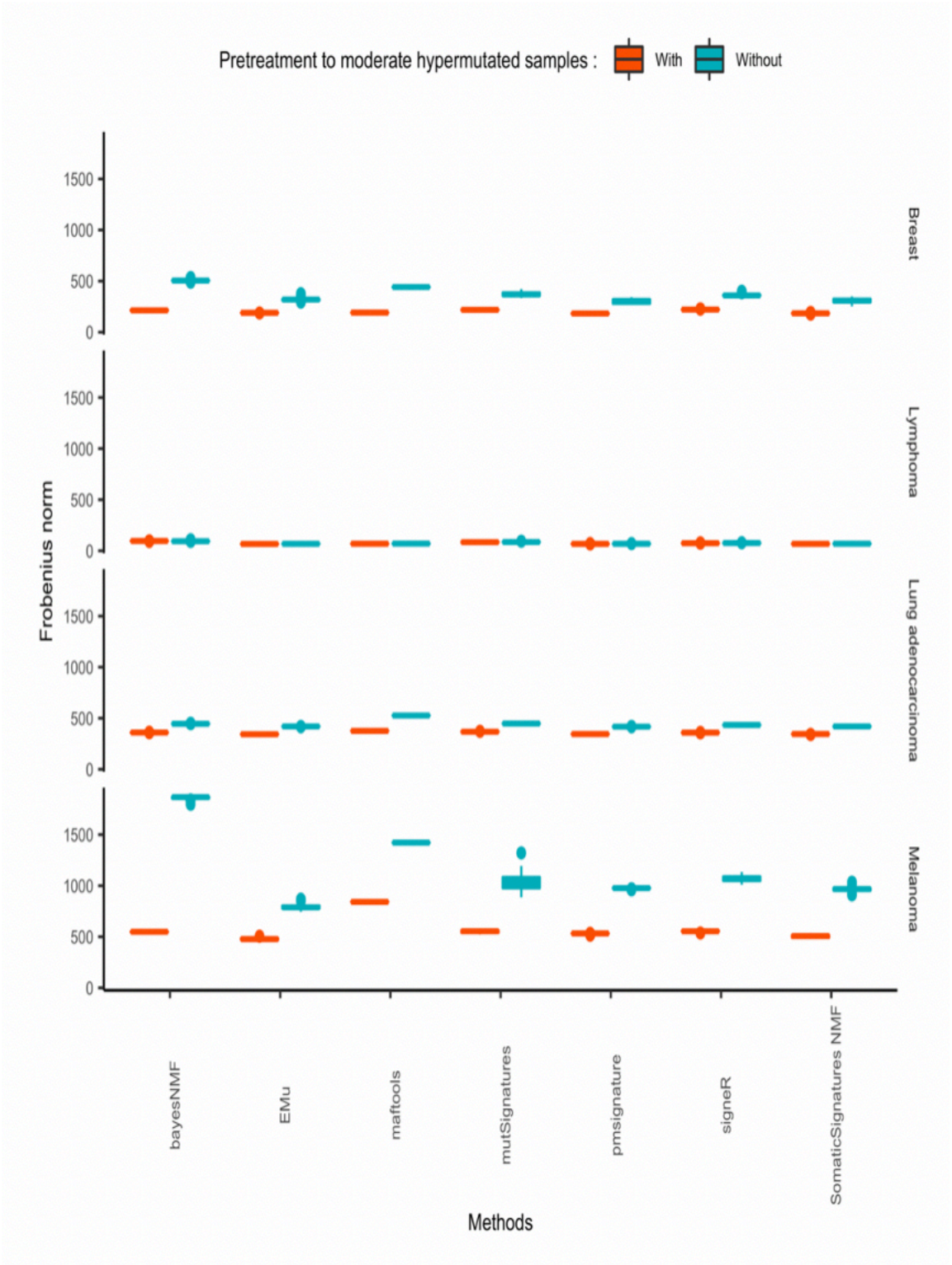


Figure III.4. **Reconstruction errors and their variability due to stochastic steps in the algorithms with and without pre-treatment to moderate the effect of hypermutated samples.**

5.1.2 CONFUSION MATRICES

Realistic simulations were used to evaluate the performance of each method for de novo extraction followed by a classification step in which the extracted signatures are assigned to the most similar COSMIC signature.

Figures III.5 and III.6, respectively show the specificity and sensitivity of such two-stage procedure as functions of the number of samples G in each catalogue and the cosine similarity cut-off h, while Figures III.7 and III.8 show the specificity and sensitivity as functions of the number of mutations in each catalogue and h.

We do not see very large differences in the tools' specificity with respect to the number of samples (Figure III.5). For Profiles 2, 3 and 4 the specificity of all methods is close to 1 even for small sample sizes, while for Profile 1, that is characterised by small contributions from several signatures (Figure 25), the specificity is close to 1 starting from 50 samples. The sensitivity of most of the algorithms increases with the sample size (Figure III.6) and this trend is more evident for Profile 1. Methods based on NMF (maftools, SomaticSignatures, mutSignatures) have lower sensitivity, while methods based on probabilistic models perform better, with the notable exception of signeR. Most of the differences between tools are observed for Profile 4, with some methods (Emu, bayesNMF, pmisignature) having a sensitivity close to 1 and the others having lower and more variable sensitivities.

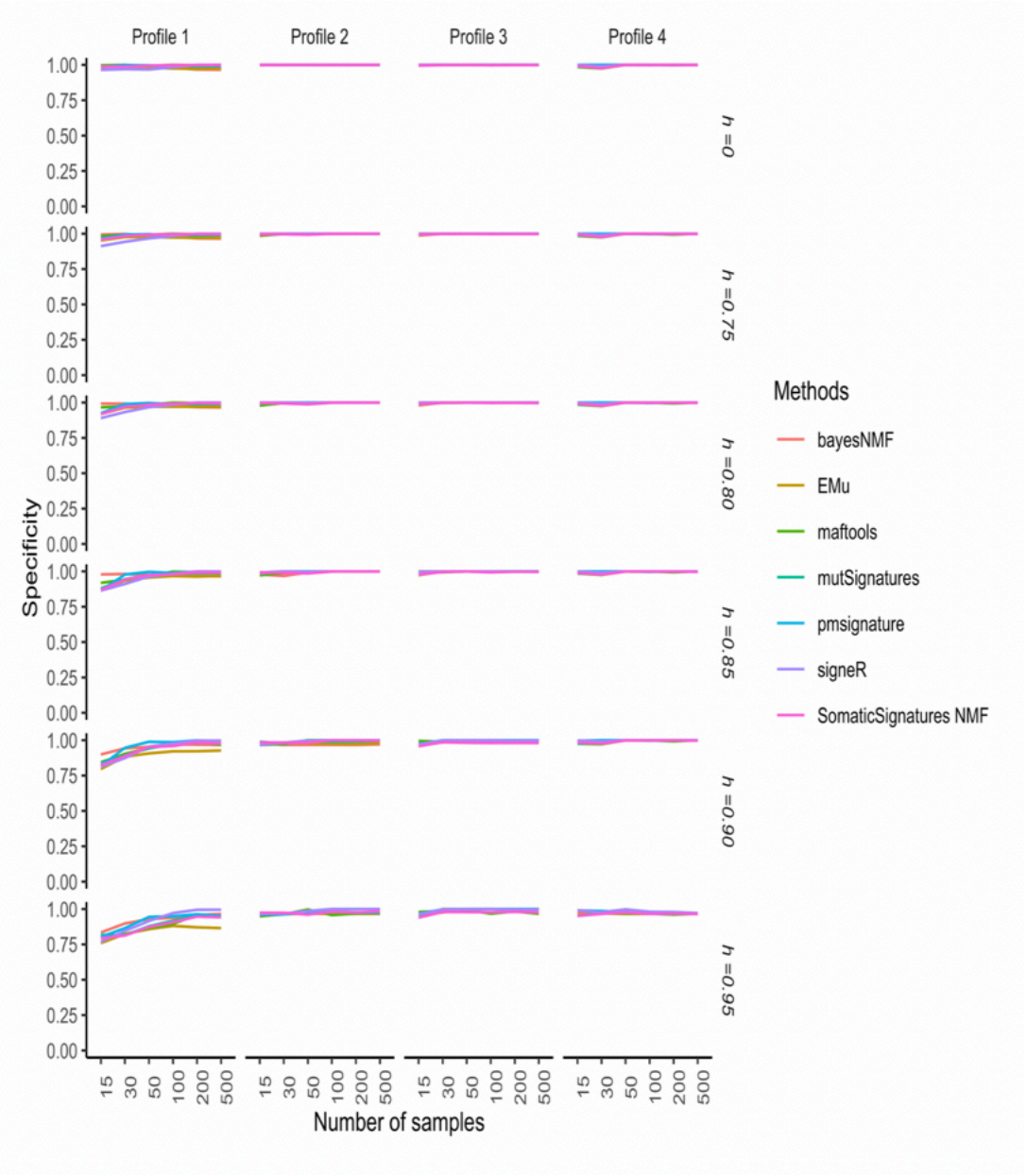


Figure III.5. **Simulation study: specificity of extraction methods and mapping on COSMIC signatures as the number of analyzed catalogues and the cosine cut-off *h* vary.**

Specificity is estimated from 50 replicates each made of G genomes. The average number of mutations in each catalogue is $r = 10{,}000$. The model used to simulate realistic replicates according to the four Profiles and the estimation methods are described in the section Data and experimental settings.

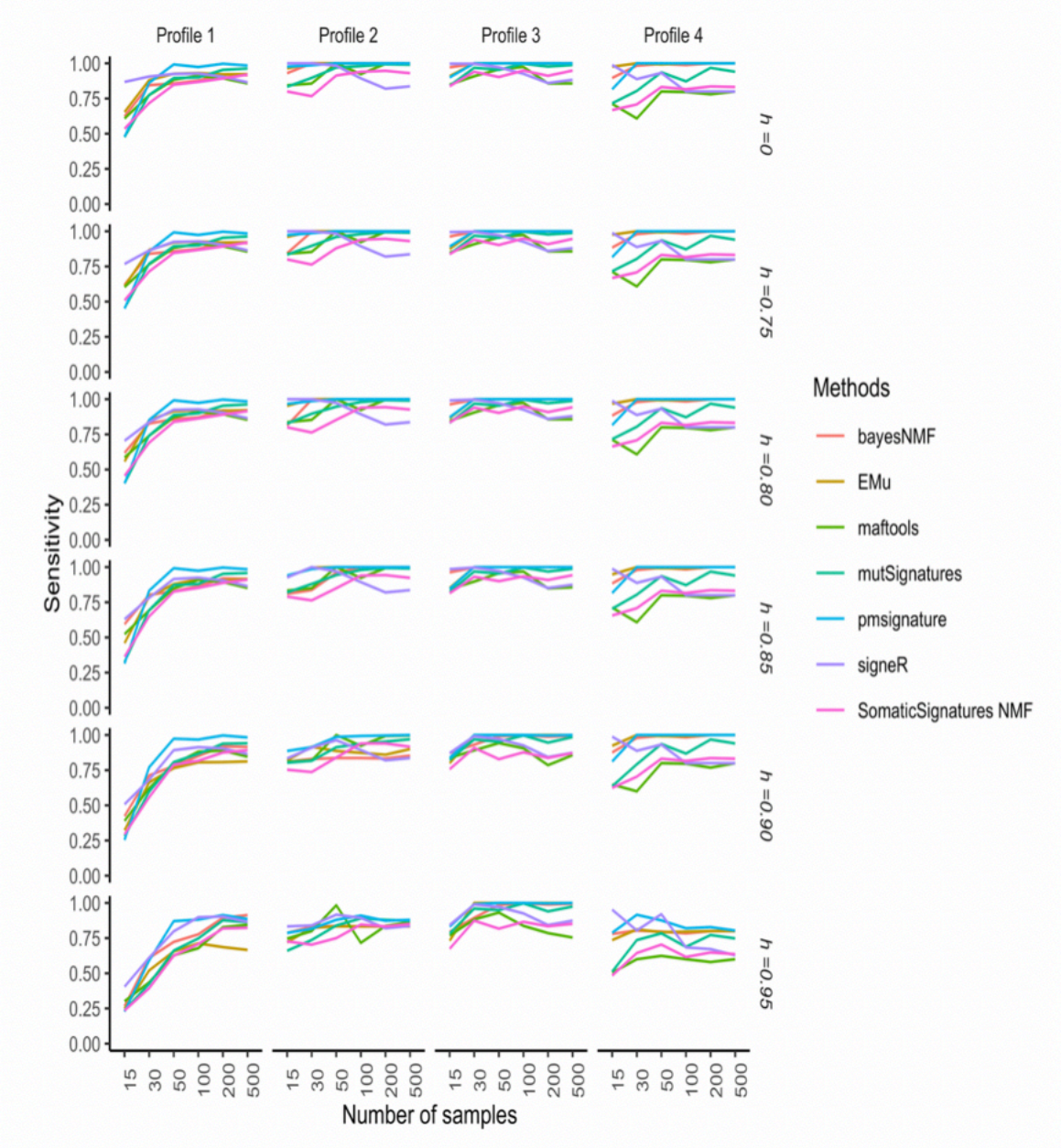


Figure III.6. **Simulation study: sensitivity of extraction methods and mapping on COSMIC signatures as the number of analyzed catalogues and the cosine cut-off *h* vary.**

Sensitivity is estimated from 50 replicates each made of G genomes. The average number of mutations in each catalogue is $r = 10{,}000$. The model used to simulate realistic replicates according to the four Profiles and the estimation methods are described in the section Data and experimental settings.

Specificity increases with the average number of mutations (Figure III.7). For Profiles 2,3 and 4 it is close to 1 starting from as low as 1000 mutations, while for Profile 1 it is only for at least 10,000 mutations that we observe a specificity close to 1 for most of the methods, with the notable exception of bayesNMF that performs well even for lower numbers of mutations. Sensitivity increases with the average number of mutations with a large variability according to the cancer profile and method (Figure III.8). Sensitivity is high for cancer profiles characterized by one predominant signature (Profiles 3 and 4) or two strong signatures (Profile 2) but may become relatively low for datasets characterized by small contributions by several signatures (Profiles 1). This indicates that signatures that act together with other signatures and have small effects may be more difficult to identify.

Specificity and sensitivity slightly deteriorate for higher cut-off values. This is expected because by setting a higher cut-off, the number of found signatures that are not similar enough to COSMIC signatures increases. Because these estimated signatures are considered as novel, they are false positives (that is found signatures not used for simulations), leading to a greater number of false positives and therefore to a lower specificity. Moreover, if the cut-off is too stringent, the number of false negatives will be high because some signatures used for the simulations are correctly found but do not score a high enough cosine similarity and therefore count as false negatives. This will make the resulting sensitivity low.

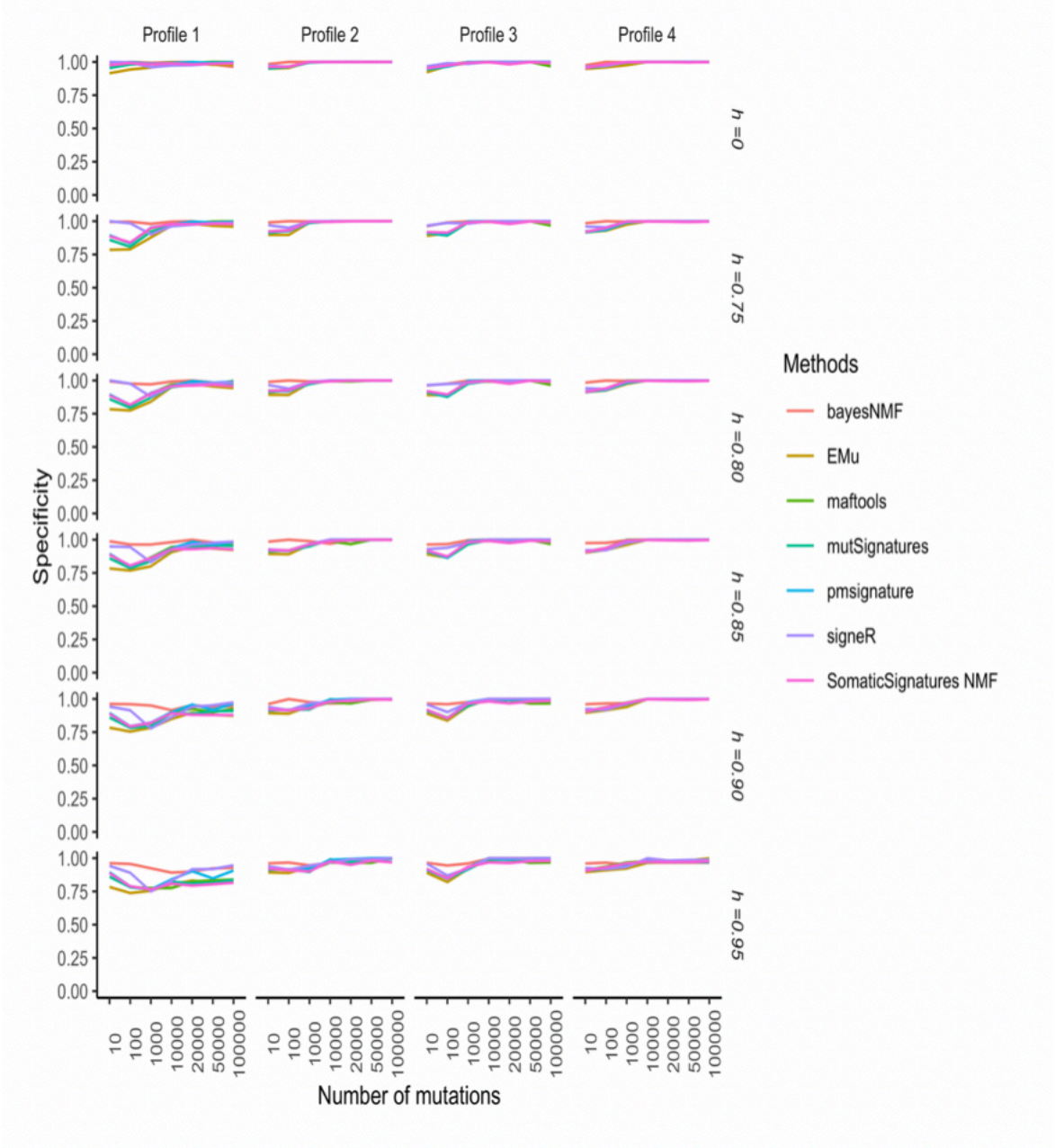


Figure III.7. **Simulation study: specificity of extraction methods and mapping on COSMIC signatures as the average number of mutations and the cosine cut-off *h* vary.**

Specificity is estimated from 50 replicates each made of $G = 30$ catalogues. The model used to simulate realistic replicates according to the four Profiles and the estimation methods are described in the section Data and experimental settings.

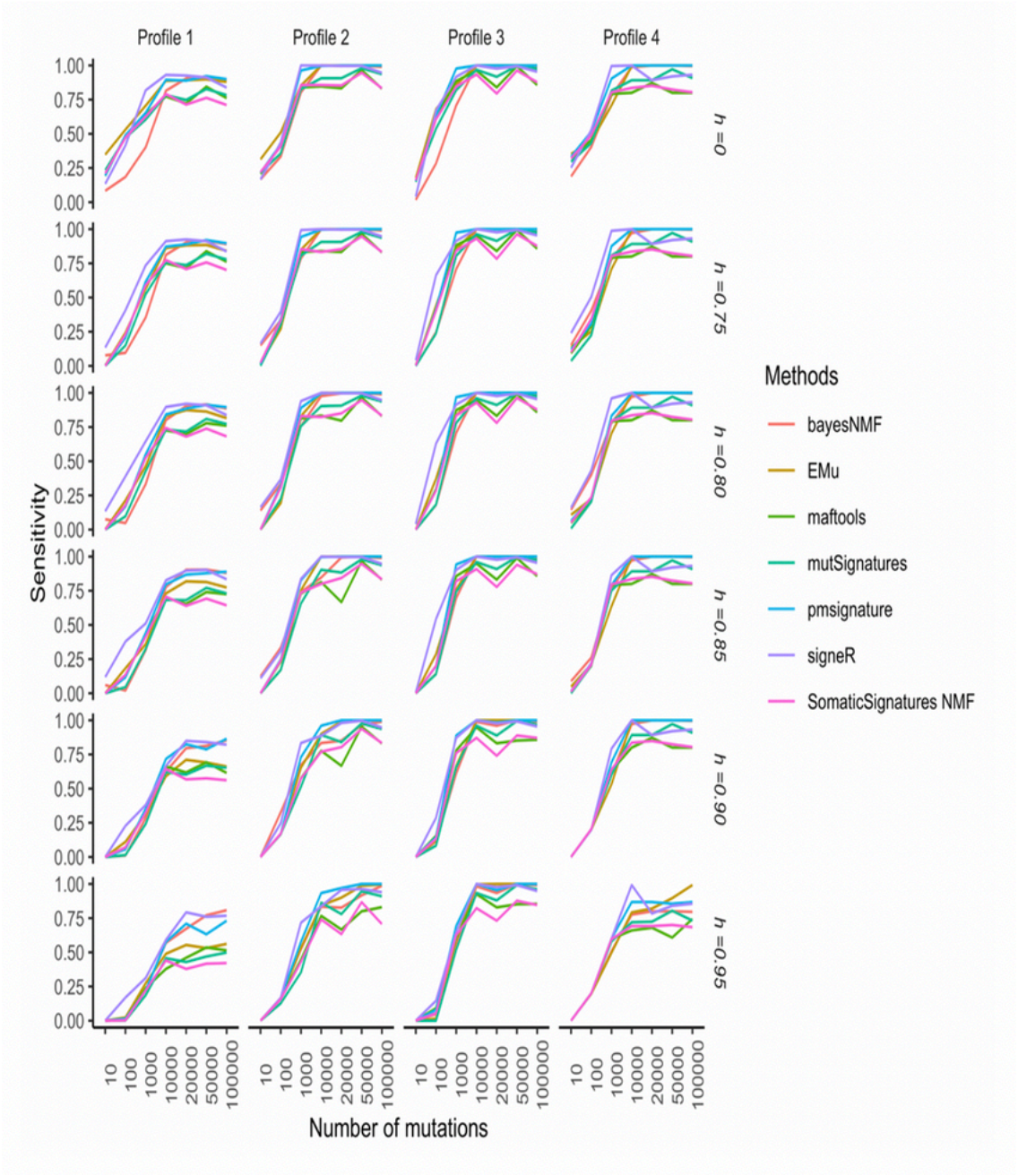


Figure III.8. **Simulation study: sensitivity of extraction methods and mapping on COSMIC signatures as the average number of mutations and the cosine cut-off *h* vary.**

Sensitivity is estimated from 50 replicates each made of $G = 30$ catalogues. The model used to simulate realistic replicates according to the four Profiles and the estimation methods are described in the section Data and experimental settings.

Methods were also evaluated with regards to running time. Figure III.9 show the running time when tools are applied to real lung datasets with a varying number of samples. While all methods show a fast-growing running time with increasing number of samples, SomaticSignatures and maftools are much faster than the others for more than 100 samples, making it possible to analyse large number of samples in few seconds. For example, for two hundred samples, the slowest method (signeR), the running time is 913.72s while for the fastest (maftools), the value is 5.97s.

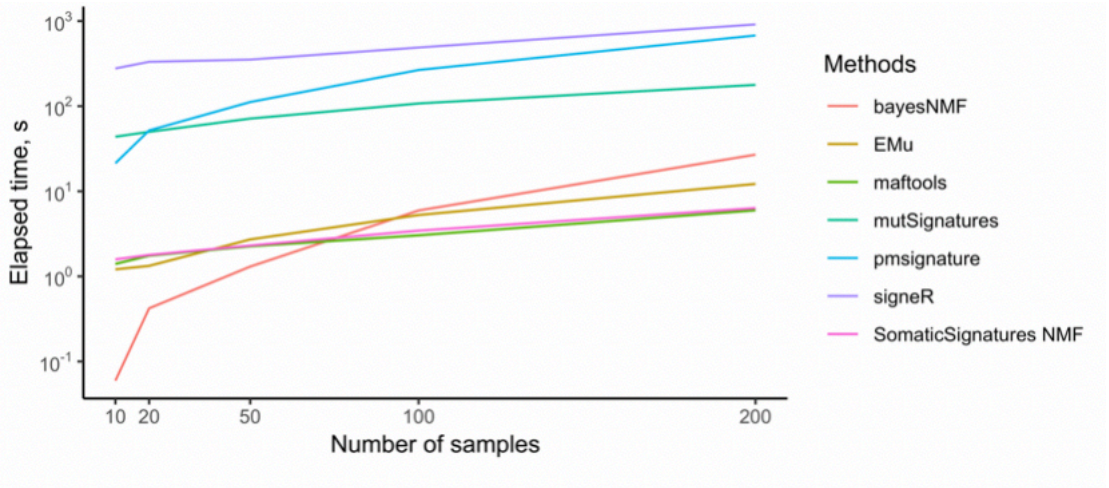


Figure III.9. **Running times of *de novo* tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes.**

The y-axis is in logarithmic scale.

5.2 PERFORMANCE OF REFITTING TOOLS

The distribution of the differences between the estimated and true contributions of all n signatures $e_g^1, \ldots, e_g^{30}$ for the different refitting methods under evaluation is shown in Figure III.10. Sample catalogues were simulated mimicking Lung cancer profiles (Profile 3), with signatures 1,2,4,5,6, 13 and 17 actually contributing as shown in Figure III.1. All methods give almost identical results.

By comparison with the true exposure profile given in Figure III.10, it is clear that all refitting methods provide good estimates of the contributions of all but signatures 4,5 and, 17 and to a lesser extent signature 6. Moreover, all methods correctly estimate a zero contribution for signatures 3 and 16 even though these are very similar to signature 5, Figure I.7.

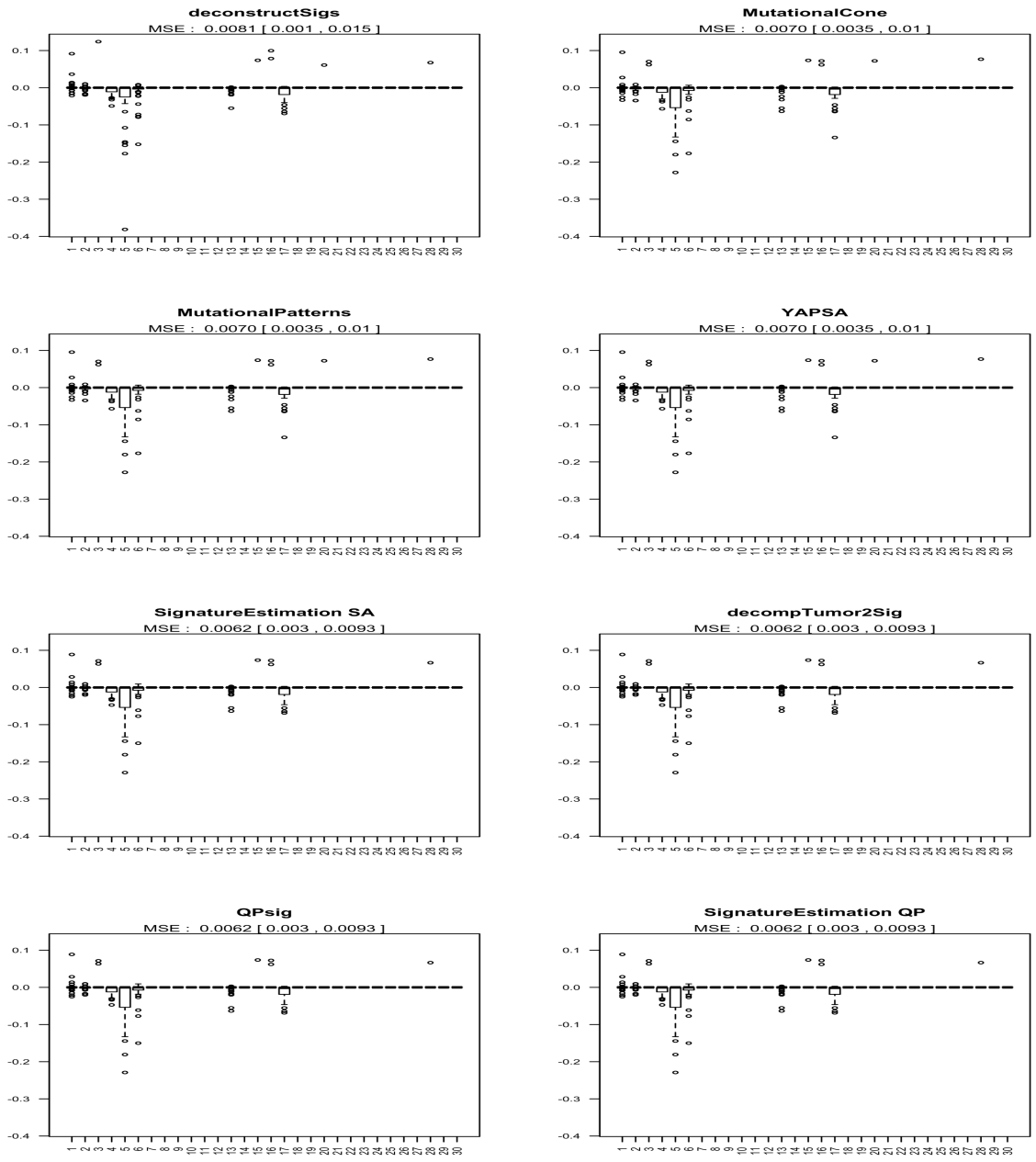


Figure III.10. **Simulation study: bias of the estimates of each signature contribution for several refitting methods.**

For each signature, the bias estimates are obtained by averaging the exposure estimates across 50 samples. Mean square errors, together with 95% confidence intervals, are reported on the top of each plot. Simulations were done according to the model described in the Data and experimental settings section.
Interestingly, signatures 2 and 13 (both attributed to APOBEC activity) are in general well identified by all methods. This finding is in line with previous claims about the stability of these two signatures.

In terms of running time, deconstructSigs and SignatureEstimation based on simulated annealing are more than two orders of magnitude slower than the other methods (Figure III.11). All other methods run in a fraction of second. As expected, the running time increases linearly with the number of samples. MutationalCone, our custom implementation of the solution to the optimization problem solved by YAPSA and MutationalPatterns outperforms all other methods. The second fastest method is SignatureEstimation based on Quadratic Programing. As example, for two hundred samples, the execution time of deconstrucSigs is 86.148s and for MutationalCone is 0.028s.

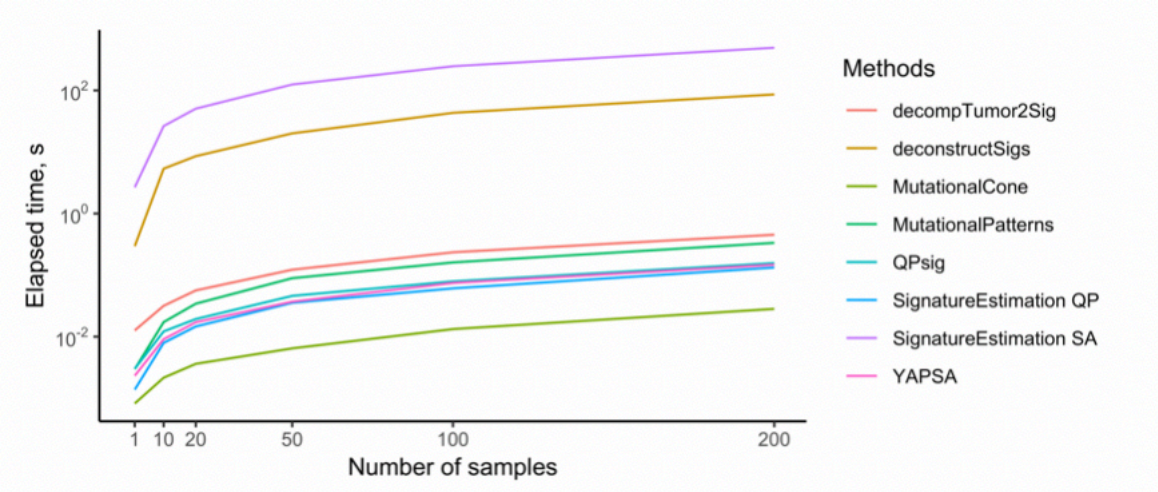


Figure III.11. **Running times of refitting tools. Methods were applied to subsets of the TCGA Lung cohort of different sizes.**

The y-axis is in logarithmic scale.

6. CONCLUSION

In this work, we complement and expand a recent review of the available methods to identify mutational signatures[173] and we compare their performance using both real world TCGA data and simulated data. The results of the work presented in this chapter can lead to a better understanding of the strengths and limitations of each method as well as to the identification of the key parameters influencing their performance, namely the number of mutations and the "complexity" of the contributing signatures.

We have demonstrated that it is mainly sensitivity and not as much specificity that significantly decreases when underlying signatures are more "complex". An intuitive reason for this result is that a signature with low impact is difficult to detect and therefore will be wrongly considered as a "negative"; several such signatures will then imply a large number of false negatives, i.e. low sensitivity. Indeed, recent evidence shows that the majority of cancers harbor a large number of mutational signatures [21] and therefore belong to the latter scenario.

With regards to the mutation number, we observe that with the number of mutations that could be found in some cancer exomes the performance is generally poor (i.e. low specificity and sensitivity). This problem is likely to be mitigated if counts were normalized by the expected number of each type's trinucleotides in the analyzed region under healthy condition, that is if an opportunity matrix was provided. We do not address this important aspect in our comparison study as only a few methods can incorporate opportunity matrices.

Additionally, we showed that when comparing identified signatures with COSMIC signatures, the choice of a cosine similarity cut-off has a relatively small impact on the overall performance. If the aim is to identify novel signatures it would be preferable to choose a lower value (0.75 or less). On the contrary, if the aim is to assess the presence of known signatures in mutational catalogues (cancer genomes or exomes), we recommend turning to refitting methods. For well-studied cancers, refitting approaches are a faster and more powerful alternative to de novo methods, even with just one input sample. As the COSMIC database has been built and validated by analyzing tens of thousands of sequences of most cancer types, we recommend borrowing strength from previous studies and using refitting tools when performing standard analysis not aimed at the discovery of new signatures.

Our simulation study seems to indicate that de novo probabilistic methods EMu and bayesNMF have an overall better performance as they achieve better sensitivity and specificity with a fair running time. However, in order to assess the robustness of new results, due to the variability of outcomes and the

presence of hypermutated samples, we recommend to systematically perform a sensitivity analysis based on the application of one or more alternative methods based on different algorithms.

Our analysis also reveals that if the dataset under consideration contains catalogues with a very large number of mutations, all methods achieve better performance by replacing such outliers with the bayesNMF pre-treatment function get. lego96.hyper. Interestingly, the mutation profiles of the synthetic datasets simulated with our ZIP model resemble the profiles of datasets after such pre-treatment

Not all the de novo methods we evaluated offer the possibility to automatically choose the number of signatures to be found. For instance, the popular SomaticSignatures only provides a graphical visualization of the residual sum of squares for several choices of the number of signatures; the user can choose the optimal number by identifying the inflexion point. For this reason, we did not address this crucial aspect in our empirical assessment. Similarly, we only considered mutation types defined by the trinucleotide motifs, as currently only pmsignature[153] can consider more than one flanking base on each side of the substitution.

Finally, we introduced a new simulation model based on the zero-inflated Poisson distribution that allows for sparse contribution of signatures and thus makes it possible to build mutation count data that are more realistic than the pure Poisson model previously considered[13,151].

CHAPTER IV:

ASSOCIATION BETWEEN PERSISTENT ORGANIC POLLUTANTS AND DNA METHYLATION

As previously described in Chapter I, PFASs and BFRs have been classified as POPs for their tendency to be extremely stable and persistent in the environment, having long half-lives in soils, sediments, air, or biota[61]. Human exposure to PFASs and BFRs is mainly attributable to the diet and in particular to foods of animal origin. Overall, diet accounts for over 90% of a person's POPs body burden and Human Biomonitoring (HBM) studies have revealed that PFASs and BFRs are ubiquitously present in the blood of populations of Western countries[174,175].

Emerging evidence suggests that exposure to EDCs can influence epigenetic changes such as DNA methylation. However, this evidence is mainly based on studies of exposure to compounds such as phthalates or bisphenol A, and very few studies are available on the epigenetic effects of exposure to PFASs and BFRs. In addition, most of them investigated effects on global DNA methylation while studies focusing on specific genomic regions and single CpGs are lacking.

In this chapter, using data from a French prospective cohort, we aimed to determine in which way DNA methylation could be used as a biomarker of exposure to BFRs or PFASs. For each pollutant, estimation from dietary exposure and measure of circulating levels in blood were explored.

Contribution

First author, discussed the analytical strategy with the supervisors, conducted statistical analyses and wrote the first drafts of the manuscripts.

1. Materials: the E3N prospective cohort 119

1.1 Presentation of the cohort 119

1.2 Epidemiological data collected in E3N 119

1.3 Measurement of circulating levels of BFRs and PFASs 124

1.4 Assessing DNA methylation in E3N 125

2. Statistical analyses 127

2.1. Descriptive statistics 127

2.2 Association measures 127

2.3 Gene Set Enrichment Analysis 131

3. Methylation signatures of Brominated Flame Retardants 133

3.1 Approaches 133

3.2 Findings 134

4. Methylation signatures of Per- and polyfluorinated Alkylated Substances 146

4.1 Approaches 146

4.2 Findings 147

5. Conclusion 155

5.1 Methylation signatures of Brominated Flame Retardants 155

5.2 Methylation signatures of Per- and polyfuorinated alkylated substances 156

1. MATERIALS: THE E3N PROSPECTIVE COHORT

1.1 PRESENTATION OF THE COHORT

E3N, the Étude Épidémiologique auprès de femmes de la *Mutuelle Générale de l'Éducation Nationale* (MGEN) is an ongoing French prospective cohort study investigating risk factors (lifestyle, nutritional, hormonal and genetics) associated with health outcomes (cancer or non-communicable diseases) in women. This cohort is run by the INSERM (National Institute for Health and Medical Research) "Health across generations" team at the Gustave Roussy Institute in Villejuif, France.

E3N started in 1990 and involves around 98.995 French women born between 1925 and 1950, who were living in metropolitan France at inclusion and were insured by the MGEN, a national health insurance scheme for workers in the French education system, a large part of whom are teachers. At the time of its creation in 1990, it was the largest epidemiological cohort study in France. In 1993 it joined other European cohorts to establish the European Prospective Investigation into Cancer and Nutrition (EPIC) study, a consortium of prospective cohort studies coordinated by the International Agency for Research on Cancer (IARC) of which E3N became the French component. The aim of EPIC is to investigate the relationship between the diet, lifestyle, nutritional and metabolic characteristics on cancer and other chronic diseases.

Initiated by Dr Françoise Clavel-Chapelon, the E3N study received ethical approval from the The French National Commission for Computed Data and Individual Freedom (Commission Nationale Informatique et Libertés, CNIL).

1.2 EPIDEMIOLOGICAL DATA COLLECTED IN E3N

1.2.1 DATA COLLECTION

During the inclusion phase to establish the cohort, between January 1989 and 1990, 500 000 women were invited to join in the study and 20% of them agreed to participate by signing an informed consent and completing a baseline questionnaire. The date of completion of the baseline questionnaire as indicated by the participants was considered as the date of recruitment.

Since then the cohort has been followed-up through self-administered questionnaires sent approximately every two years. Up to 2018, twelve questionnaires have been sent to the E3N women (Figure IV.1).
The questionnaires include questions on anthropometry (e.g. weight, height, waist circumference), lifestyle (e.g. tobacco and alcohol consumption), socio-demographic factors (educational level, profession), hormonal factors (e.g. age at menarche and at menopause, use of hormone replacement

therapy or oral contraceptives), reproductive factors (e.g. age at first birth and parity), family history of cancer, use of various medications as well as questions on personal history of various diseases (e.g. cancer, myocardial infarction, stroke and others). Several questions on menopause, anthropometry and tobacco smoking, and about the diagnosis of cancer and other diseases were repeated for each questionnaire.

Between 1994 and 1999, a biological bank was created with the collection of blood samples donated by approximately 25 000 E3N participants while between 2009 and 2011, about 47 000 saliva samples were further collected from women who had not donated blood samples in order to have the possibility to perform genotyping of around three quarters of the entire cohort

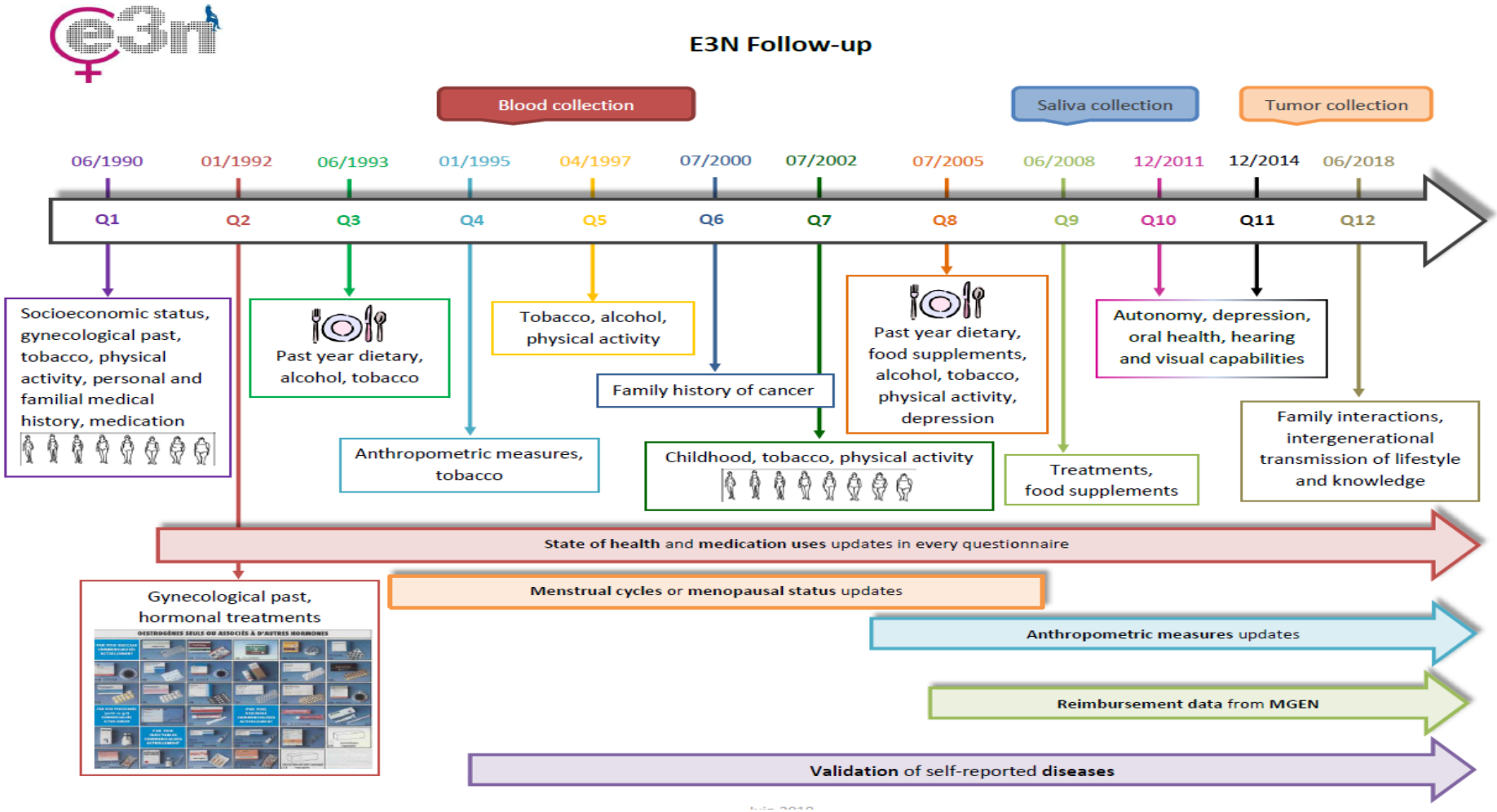


Figure IV.1. **Calendar of self-administrated questionnaires in E3N**

1.2.2 DIETARY QUESTIONNAIRE

In E3N, detailed information on dietary habits was collected twice through two extensive food frequency questionnaires for the third and eight follow-ups (Q3, in 1998 and Q8 in 2005). The two questionnaires included respectively 238 questions on frequency of consumption of specific foods, selected on the basis of the French meal pattern.

The first food frequency questionnaire (FFQ) was sent to 93 055 women and had a response of 82% (76 208) while the second was sent to 93 121 women with a response of 77% (71 788). For each FFQ, questions concerned foods and drinks across eight consumption occasions from breakfast to after-dinner snacks and were designed to assess the habitual diet of the previous year.

The questionnaire was structured into two parts with the first one related to the quantification of food consumption and the second one describing the qualitative aspects of different food items within each food group. Based on 66 food groups, the quantitative section described the habitual frequency and portion sizes consumed using an album including 42 food groups while the rest were estimated in natural units (e.g. number of eggs, tablespoons).

The second part of the questionnaire contained qualitative questions concerning food items within each food group listed in the first part of the questionnaire with study subjects asked to score their relative consumption frequency (never, 1-3 times/month or 1-7 times/week) for each food item within the group.

1.2.3 THE E3N-TDS2 DATABASE ON INDIVIDUAL EXPOSURE TO CONTAMINANTS

The Second French Total Diet Study (TDS2), conducted in 2006 by the French Agency for Food, Environmental, and Occupational Health (ANSES), assessed exposure to more than 400 contaminants in a large number of foods representative of the French diet in order to assess the risks of exposure to chemical substances in relation to public health. Another main objective of the study was to provide scientific information that would enable authorities to control and regulate chemical products and the safety of food product[176]. Briefly, data on consumption trends and eating habits from the second French individual food consumption survey (INCA2) as well as data from a 2004 purchase panel of French households (SECODIP) were used to identify the core foods to be sampled.
Finally, 186 core foods on a national scale and 70 core foods on a regional scale were selected according to (1) consumption data for adults and children, (2) their consumer rates, and (3) contribution to exposure to one or more contaminants of interest[176].

Thus, between 2007 and 2009, in eight greater regions of the French metropolitan territory, a total of 20 280 different food products were purchased to make up the 1352 composite samples of core foods to be

analyzed for additives, environmental contaminants, pesticide residues, trace elements and minerals, mycotoxins and acrylamide. A total of 445 different chemical were analyzed in the food samples and results of the study are publicly available online (data.gouv.fr).

To estimate the individual dietary exposure to chemical substances for each E3N participant, food items reported in the TDS2 study have been matched to those of the E3N food questionnaire leading to the E3N-TDS2 database (Mancini and colleagues, *paper in progress*).
Individual estimates of dietary exposure to BFRs and PFASs for each E3N participant were available from previous work coordinated by Francesca Mancini[76] in the context of research programs on type 2 diabetes and hormone-related cancer (e.g. project ED-Cancer funded by INSERM Plan Cancer). In brief, estimates of consumption of each food item obtained through the first E3N food frequency questionnaire were coupled with data from the ANSES survey of levels of BFRs and PFAS measured in the corresponding food item.

Estimation of dietary exposures to BFRs and PFASs in E3N cohort was based on data from the dietary questionnaire completed by E3N participants in 1993. The validity and reproducibility of the questionnaire have been previously described by van Liere and colleagues[177] and was designed to estimate food consumption over the previous year for a set of 238 food items consumed on eight occasions from breakfast to dinner snack.

Through the merging of the E3N food frequency questionnaire and the TDS2 contamination database a E3N-TDS2 database has been created which allowed to estimate the individual dietary exposure to HBCDs congeners (HBCDalpha, HBCDbeta and HBCDgamma), PBDEs congeners (BDE-47, BDE-99, BDE-100, BDE-153, BDE-154, BDE-183 and BDE-209), PFOA and PFOS for each woman in E3N cohort.
For food items with values of contamination below the Limit Of Detection (LOD), a value of ½ LOD was assigned and exposure estimates used for our analyses is expressed in ng/kg body weight (BW)/day.

1.3 MEASUREMENT OF CIRCULATING LEVELS OF BFRS AND PFASS

In addition to the estimates of dietary exposure to BFRs and PFASs obtained for all E3N cohort participants that completed the food frequency questionnaire, circulating levels of BFRs and PFAS were measured in a case-control study of 200 breast cancer cases and 200 controls nested within E3N using blood samples.

1.3.1 DESIGN OF THE CASE-CONTROL STUDY

For the nested case-control study on breast cancer, only women that provided blood samples, filled the dietary questionnaire (Q3), and participated in the follow-up after blood collection were considered. Those with any type of prevalent cancer at Q3 and missing values for matching criteria (such as age, BMI, menopausal status) were excluded. Women diagnosed with breast cancer (both in situ or invasive) after 1993 (Q3) and up to the end of 2014 (Q11) who donated a blood sample were considered as cases. Controls were selected from women without a diagnosis of cancer at the date of diagnosis of the corresponding case.

Finally, a total of 197 case-control pairs nested within the E3N cohort were matched on age at blood collection (± 2 and 3 years) , BMI (< vs. ≥ 25kg/m²), menopausal status, date (± 3 months) and department of residence at blood collection (grouping of 75, 77, 78, 91, 92, 93, 94, 95 / 10, 89 / 01, 73, 74 / 42, 43 / 27, 76 / 02, 60 / 13, 30, 84).

1.3.2 CIRCULATING LEVELS OF BFRS

Circulating levels of BFRs for the 197 breast cancer cases and 197 controls in E3N have been measured in plasma samples by the LABERCA laboratory (Oniris Nantes, FRANCE). Methodologies applied to isolate, detect, and quantify the PBDE congeners (BDE-28, BDE-47, BDE-99, BDE-100, BDE-153, BDE-154) and PBB-153 have been described by Cariou and colleagues[178]. In summary, plasma samples were first submitted to a liquid/liquid extraction with pentane and the resulting extracts were weighed to measure fat content using an enzymatic method (Biolabo; Maizy, France) before reconstitution in hexane for further purification. Then, determinations were performed using gas chromatography (Agilent 7890A) coupled to high-resolution mass spectrometry (GC-HRMS) on double sector instruments (JEOL MS 700D and 800D) after electron impact ionization (70 eV), operating at 10 000 resolutions (10% valley) and in the single ion monitoring (SIM) acquisition mode. Finally, as describe by Akins and colleagues, the total plasma lipid (TPL) levels were calculated by combining the concentration of phospholipids (PHO), triacylglycerides (TAG), total cholesterol (t.CHO) and free cholesterol (f.CHO) as follows: TPL=1.677*(t.CHO-f.CHO)+f.CHO+TAG+PHO)[179].
All the analyses have been conducted in an ISO 17025:2005 accredited laboratory.

For BDE-47, BDE-99, BDE-100, BDE-153, PBB-153, all samples have been quantified – i.e. none was below the LOD. For those samples for which levels were below the LOD (1 sample for BDE-28 and 99 samples for BDE-154) the measure has been replaced by ½ LOD.

1.3.3 CIRCULATING LEVELS OF PFASS

Circulating levels of PFOA and PFOS for 388 women in the breast cancer case-control study nested in E3N have been measured in serum samples using liquid chromatography coupled to tandem mass spectrometry (LC–MS/MS) as detailed in a previous publication[94].
Briefly, the quantification was achieved according to the isotopic dilution method (i.e., using ^{13}C labeled analogous as internal standards) and the lipid content was determined with enzymatic kits (Biolabo, Maizy, France) independently for phospholipids (PL), triglycerides (TG), total cholesterol (TC) and free cholesterol (FC). Total serum lipids (TSL) were estimated using the Akins and colleague's formula as described in the previous section.
All the protocol wad based on a fully validated (2002/657/CE decision) and accredited methods (ISO 17025 standard) and all samples had levels above the LOD.

1.4 ASSESSING DNA METHYLATION IN E3N

In the same breast cancer case-control study in which circulating levels of BFRs and PFAS were measured, the Illumina® Infinium HumanMethylation EPIC array on DNA extracted from buffy coat samples were used to assess DNA methylation at more than 850 000 CpG sites across the genome.
DNA extraction, bisulfite conversion of the extracted DNA, quality control analyses, the running of the methylation assays as well as the methylation data pre-processing were performed at the Italian Institute of Genomic Medicine (IIGM) in Turin, Italy according to manufacturers' protocols and procedures developed by IIGM for previous studies on DNA methylation[180,181].

Genomic DNA was extracted from buffy coats using the QIAsymphony DNA Midi Kit (Qiagen, Hilden, Germany). Five hundred nanograms (1 microgram for a few samples) of DNA were bisulphite-converted using the EZ-96 DNA Methylation-Gold™ Kit (Zymo, California, USA) and hybridized to Infinium Human Methylation EPIC BeadChips (Illumina, California, USA). Each chip was subsequently scanned using the Illumina HiScanSQ system, and sample quality was assessed using control probes on the microarrays. Raw intensity data were finally exported from Illumina GenomeStudio (version 2011.1). Samples were distributed into 96-well plates and processed in chips of 12 arrays (8 chips per plate) with case–control pairs arranged randomly on the same chip.

Data pre-processing was carried out using an in-house software written for the R statistical computing environment[180].

For each sample and each probe, measurements were set to missing if obtained by averaging intensities over less than three beads, or if averaged intensities were below detection thresholds estimated from negative control probes. Background subtraction (to remove background noise) and dye bias correction (for probes using the Infinium II design) were also performed. The resulting subset of 867 867 CpG loci was selected for further analyses, and among these, probes with missing values in more than 5% of the samples were excluded from the analyses, leaving 805 837 probes. Samples with more than 5% of non-detected probes were also excluded from the analysis. The final dataset included one hundred and sixty-eight case-control pairs the passed the pre-processing step for which and included methylation measures for 805 837 CpGs.

2. STATISTICAL ANALYSES

Statistical analyses were based on the objectives described in Chapter I and were performed using the R 3.5.X software.

2.1. DESCRIPTIVE STATISTICS

2.1.1 MEDIAN, FREQUENCY AND OTHER BASICS STATISTICS

For the description of the study samples, basic statistics were used such as frequency, mean, standard deviation (SD), and median value. In all analyses presented in this chapter, independent variables (e.g. levels of exposure to BFRs and PFAS) were categorical and chi-square tests were used in order to compare some characteristics of the participants, which helped identify potential confounding factors to be considered in further analyses.

2.1.2 QUANTILE-QUANTILE PLOT

The quantile-quantile (Q-Q) plot is graphical technique generally used to determine if two data sets come from populations with a common distribution. It is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, the points will form a line that's roughly straight.

2.2 ASSOCIATION MEASURES

Linear mixed models (LME) are an extension of the simple linear model to allow for the inclusion of both fixed and random effects that contribute linearly to the response function. Such models are particularly useful when there is non-independence in the data as it is the case for the DNA methylation measures that may vary according to technical factors such as chip and plate that are hierarchically organized (Figure IV.2).

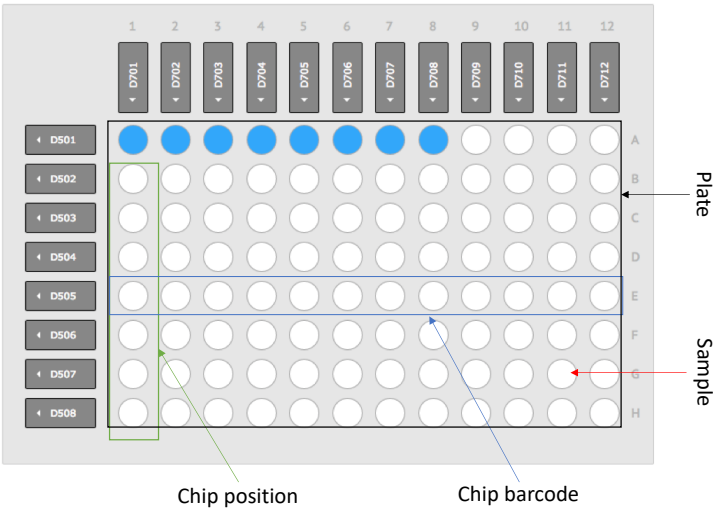


Figure IV.2. **Organization of chips within plate**

Lme have been widely used in many research areas, especially in the area of psychometrics, sociology and biomedical research, to analyze longitudinal and clustered data. Like other statistical models, these models describe a relationship between a response variable and some regressors that have been measured or observed along with the response.

2.2.1 FIXED VS. RANDOM EFFECTS

The core of mixed models is that they incorporate both fixed and random effects. Fixed effects are variables that we are particularly interested in as we expect they will have an effect on the dependent/response variable. In our case, we are interested in making conclusions about whether POPs are associated with DNA methylation and therefore POPs will be considered as fixed effect variables.

Random effects are usually grouping factors for which we are trying to control as we know they may impact on the outcome but in which, in general, we are not particularly interested.

For example, for the methylation analyses in our study, DNA samples were placed on four plates and, as expected, the measured levels of methylation appear to be quite different across plates, especially between plates 1-2 and 3-4. Variation across plates is often observed in studies based on methylation arrays for various reasons (e.g. in our study for the last two plates 1µg of DNA was used instead of 500ng). However, since beta-values are a ratio between the methylated signal and the total signal this is unlikely to influence the results. Plate is therefore considered as a random effect, and as they may contain up to 96 samples in a same experiment, the sample position within the plate is also considered as a nested random effect.

Indeed, different random effects can be crossed or nested according to their relationship. For example, if the observations are grouped by a factor g2, which is nested within another factor g1, then the third formula in Table IV.1 can be used to model variation in the intercept with the lme4 R package, while if the data are grouped by fully crossing two factors, g1 and g2, then the fourth formula in Table IV.1 may be used.

Table IV.1. **Examples of random effects mixed-effects model formulas used in the lme4 R package.** *Adopted from Bates and colleagues*[182]

The names of grouping factors are denoted *g*, *g1*, and *g2*, and covariates and a priori known offsets as *x* and *o*.

Formula	Alternative	Meaning
(1 | g)	1 + (1 | g)	Random intercept with fixed mean
0 + offset(o) + (1 | g)	-1 + offset(o) + (1 | g)	Random intercept with a priori means
(1 | g1/g2)	(1 | g1) +(1 | g1: g2)	Intercept varying among g1 and g2 within g1
(1 | g1) + (1 | g2)	1 + (1 | g1) + (1 | g2)	Intercept varying among g1 and g2
x + (x | g)	1 + x + (1 + x | g)	Correlated random intercept and slope
x + (x || g)	1 + x + (1 | g) + (0 + x | g)	Uncorrelated random intercept and slope

2.2.2 MATHEMATICAL DEFINITION OF A LINEAR MIXED EFFECTS MODELS

For each probe, we considered mixed-effect models of the type

$$y_{ijk} = \alpha + \beta_1 x_{ijk}^1 + \beta_2 x_{ijk}^2 + \cdots \beta_p x_{ijk}^p + u_{jk} + u_k + \epsilon_{ijk}$$

Where

- y_{ijk} is the methylation level of individual i whose sample has been analyzed on chip j of plate k.
- x_{ijk}^1 is the EDCs level of individual ijk; similarly, $x_{ijk}^2, \ldots, x_{ijk}^p$ are the values of the other fixed effects for this individual (age, BMI, etc)
- $\beta_1, \ldots, \beta_p$ are the coefficients of the fixed effects. α is the fixed intercept.
- u_{jk} is the random intercept effect accounting for the average methylation level of chip j in plate k. u_k is the random intercept effect accounting for the average methylation level of plate k.
- ϵ_{ijk} is the random error of individual ijk

The hypothesis in the model are:

- $u_{jk} \sim \mathcal{N}(0, \tau^2)$ for each jk
- $u_j \sim \mathcal{N}(0, \eta^2)$ for each j
- $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$ for each ijk
- $u_{jk}$and $u_j$and ϵ_{ijk}uncorrelated.

We remind that because there are 4 plates, with 8 chips each, and each chip carries 12 samples, in principle we have $k \in \{1,4,\}\, j \in \{1, \dots, 8\},\ k \in \{1, \dots, 12\}$.
We fitted the model above with the formula:

DNA methylation ~ ED + Covariates (Age, BMI, etc.), random=~1|Plate/Chip

where the random term (~1|Plate/Chip) *allows us to control batch effects (source experimental variability) by means of a random intercept with two levels of clustering.*

2.2.3 STATISTICAL MODELING

For the present work, only data from the controls (women that have not been diagnosed at the date of diagnosis of the matched case) have been analyzed because they are more representative of the full cohort and to avoid selection bias due to conditioning on the case-control status (a colliding variable).

We assessed the association between dietary exposure and circulating levels of BFRs and PFAS with DNA methylation levels both at the global level, in specific genomic regions and for each CpG site independently. For each CpG, we computed β-values, that represent the ratio of the methylated probe intensity over the overall intensity (sum of methylated and unmethylated probe intensities). The M-values were then calculated as log2[β-value/(1–β-value)] and used as dependent variables in the regression model[183]. Global methylation was defined as the mean of M-values across all CpG sites across all the genome. Additionally, methylation levels were computed by genomic region defined according to the CpG position (e.g. in CpG Island/Shore or Shelf/Other and according to genomic regulatory features – i.e. in promoter regions or outside them).

Further details on the statistical methods and study populations related to the specific investigations performed will be described in the corresponding relevant sections.

2.2.4 FALSE DISCOVERY RATE

In modern omics research, tens of thousands of tests are conducted simultaneously, increasing the likelihood of obtaining false positives. Several statistical techniques have been developed to prevent this

multiple testing problem, controlling for different types of errors, including the Family Wise Error Rate (FWER) and the False Discovery Rate.

In omics association studies, the FDR, defined as the expected proportion of false positives among all rejections of the null hypothesis is often preferred over the FWER, the probability to obtain at least one false positive, because it leads to less conservative decision rules.

According to the Benjamini and Hochberg procedure[184], the FDR can be controlled at the desired level α by adjusting each test p-value as follows:

- order all p-values in ascending order $p_{(1)} \leq p_{(2)} \leq \cdots p_{(m)}$, where m is the number of tests
- define the adjusted p-value

$$p_{(i)}^{BH} = \frac{p_{(i)}}{i} \mathrm{m}$$

By rejecting all the null hypothesis from tests having adjusted p-values less than α, the FDR is controlled at the threshold α. Note that the Bonferroni correction for the control of the FWER is less conservative than the Benjami-Hochberg procedure because $p^{Bonf} = pm > p^{BH}$. We controlled the FDR at the threshold $\alpha = 0.05$ by computing BH adjusted p-values with the p. adjust function in the stats R package.

2.2.5 MISSING DATA

When data for a variable were missing in less than 5% of samples, missing values were imputed to the modal category of the variable. For missing variables collected through several questionnaires within the E3N cohort, imputation is performed using information provided in the previous questionnaire.

2.3 GENE SET ENRICHMENT ANALYSIS

2.3.1 OVERVIEW

Gene Set Enrichment Analysis[185] (GSEA) is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between a biological state (e.g. a alternative phenotype) or correlation with a quantitative "phenotype" (e.g. BFRs levels).

GSEA produces a ranked list of genes sets based on an enrichment score. The GSEA method can be summarized as follows:

➔ For each gene in the full list, the difference between its average methylation levels according to the categories of a categorical variable (e.g. case/control phenotype) is measured through appropriate test statistics (usually Kolmogorov-Smirnov-like

statistics). If the association of interest is with a continuous "phenotype" (e.g. BFR level) its correlation with the average methylation level of each gene is calculated.

- ➔ For each gene set, an enrichment score (ES) is computed by walking down the full list of genes, increasing a running-sum statistic when a gene is in the set and decreasing it for genes not in the set.

- ➔ The significance level of the ES (nominal p-value) is assessed using an empirical test based on the permutations of the "phenotype" variable. This allows to simulate the null distribution of the ES while preserving the complex structure of the methylation data.

- ➔ P-values are adjusted for multiple hypothesis testing. The ES of each gene set is first normalized to account for the size of the set, yielding a normalized enrichment score (*NES*). The proportion of false positives is controlled by calculating the FDR corresponding to each NES, the so-called Q-value (in this context, the FDR is the estimated probability that a set with a given NES represents a false discovery. Q-values are estimated using a method that improves the Benjamini-Hochberg procedure, by comparing the tails of the observed and null distributions for the *NES*).

In its standard procedure, if more than one beta-value is associated with a gene name, the median methylation is used. Differential methylation with respect to quantitative "phenotypes" is determined using Pearson correlation.

In this context, an FDR of 0.25 or 0.3 is generally used rather than the more classic 0.05. An FDR of 25% indicates that the result is likely to be valid 3 out of 4 times, which is reasonable in the setting of exploratory discovery where one is interested in finding candidate hypothesis to be further validated as a result of future research. Given the lack of coherence in most expression datasets and the relatively small number of gene sets being analyzed, using a more stringent FDR cutoff may lead you to overlook potentially significant results.

2.3.2 THE MOLECULAR SIGNATURE DATABASE

The Molecular Signatures Database[186] (MSigDB) is a collection of annotated gene sets for use with GSEA software. The last version v7.0 updated in August 2019 include 22596 gene sets in the Molecular Signatures Database (MSigDB) are divided into 8 major collections, and several sub-collections (Appendix 3).

3. METHYLATION SIGNATURES OF BROMINATED FLAME RETARDANTS

3.1 APPROACHES

3.1.1 ASSOCIATION BETWEEN DIETARY EXPOSURE TO BFRS AND DNA METHYLATION

In this analysis, women with aberrant energy intake (e.g. 1% and 99% extremes of the energy intake/energy expenditure ratio) were excluded (n=6). Basal metabolic rate (BMR), based on age, sex and weight (self-reported in kg), multiplied by 1.55 was used to estimate a woman's energy intake. Then, our final dataset for the association between dietary exposure to BFRs and methylation M-values consisted of a subset of 162 women with methylation data on 805.837 CpGs.

We explored the association between dietary exposure to BFRs and DNA methylation (PBDEs, HBCDs and each congener independently) through linear mixed-effects models with DNA methylation as dependent variable (either global methylation, or "regional" methylation or single probes), quartiles of BFRs as explanatory variable and plate and chips as random effects. Additionally, models were fitted with adjustment for age at blood collection (categorical, below or above the median), parity and total breastfeeding duration (no children or no breastfeeding, at least 1 child and ≤6 months breastfeeding, at least 1 child and >6 months breastfeeding), BMI (≤25 kg/m^2,>25 kg/m^2) and adherence scores to the healthy dietary pattern and the Western dietary pattern (as categorical variables, below or above the median) both derived from principal components analysis (PCA), as previously described by Edefonti and colleagues[187].

The exposure to BFRs estimated from food, was obtained on the basis of the dietary history of women in the cohort over the previous year through the response to the dietary questionnaires. Dietary patterns are potential confounders because they are associated with both "exposure" to BFRs (or rather the proxy used in our analyzes, which is calculated precisely from diet), and potentially methylation. For this reason, we adjust for dietary patterns.

3.1.2 ASSOCIATION BETWEEN CIRCULATING LEVELS OF BFRS AND DNA METHYLATION

For the analyses of BFRs (PBDEs congeners: BDE-47, BDE-99, BDE-100, BDE-153, BDE-154, BDE-183 and BDE-209) blood levels, that we conducted separately to the analyses of dietary exposure to BFRs, data were available for a slightly larger sample of women (N=168). For such analyses we used models similar to those used for the analyses of dietary exposure with the exception that adherence scores to the healthy dietary pattern and the Western dietary pattern were not included in the models adjusted for the covariates.

3.1.3 ENRICHMENT ANALYSIS

To determine whether any gene set or biological pathway is overrepresented in the list of genes whose DNA methylation are associated with circulating levels or dietary exposure to BFRs, we performed two separate gene set enrichment analyses[185]: (1) genes near CpG sites located in promoter region in which the association between circulating levels of BFRs and CpG site methylation levels are significant (unadjusted p-value < 5%) and (2) genes near CpG sites located in promoter region in which the association between dietary exposure to BFRs and CpG site methylation levels are significant (unadjusted p-value < 5%).

In the present study, we conducted GSEA analysis using GSEA_4.0.1 and the hallmark gene set[188] v7.0 processed in the MSigDB database, which is a collection of 50 gene sets that represent specific and well-defined biological states or processes and display coherent expression.

The enrichment would be considered 'significant' when the FDR<0.3. GSEA's parameters of "Enrichment statistic" was set to the "classic" item, and the parameter of "Metric for ranking genes" was set to "Pearson". 1000 permutations were carried out to evaluate the FDR and the p-value of the enrichment score with permutation type set to the "gene_set".

Description of gene sets identified in these analyses are available in Appendix 4.

3.2 FINDINGS

3.2.1 BASELINE CHARACTERISTICS OF THE STUDY POPULATION

The baseline characteristics of the study participants are summarized in Table IV.2. To study the association between BFRs and methylation of DNA from blood, data were available from 168 women for circulating levels of BFRs and from 162 women for the dietary exposure to these compounds. Median age of the study participants was 56.1 years and most of them had a healthy body mass index with only one quarter of them being overweight or obese. About 43% of them are nulliparous or never breastfed, 40% had at least one child but breastfed for less than 6 months and 17% had breastfed for more than 6 months.

From the detailed data from the food frequency questionnaire completed in 1993, between 2 and 5 years before the blood collection, dietary patterns were identified including a "healthy" dietary pattern and a "Western" dietary pattern[189]. In our study population around half of the women had a "healthy" diet and half adhered to a Western diet with a small overlap between the two groups.

Table IV.2. **Baseline characteristics of the study population**

	BFRs Circulating levels (N = 168)	Dietary exposure to BFRs (N = 162)
Age (%)		
<56.1	83 (49.4)	79 (48.8)
>56.1	85 (50.6)	83 (51.2)
Body Mass Index (%)		
<25	124 (73.8)	121 (74.7)
>25	44 (26.2)	41 (25.3)
Score of adherence to the healthy dietary pattern (%)		
Above median		88 (54.3)
Below median		74 (45.7)
Score of adherence to the Western dietary pattern (%)		
Above median		81 (50.0)
Below median		81 (50.0)
Parity and total breastfeeding duration (%)		
Nulliparous or never breastfeed	73 (43.5)	70 (43.2)
Parous and breastfeed for less than 6 months	68 (40.5)	65 (40.1)
Parous and breastfeed for more than 6 months	27 (16.1)	27 (16.7)

The levels of dietary exposure to BFRs estimated in our study population are presented in Table IV.3. For PBDEs congeners, the highest dietary exposure is due to BDE-47 and BDE-209 with minimum and maximum daily intakes for these congeners ranging from 0.038 to 0.445 ng/kg BW/day and from 0.1 to 0.823 ng/kg BW/day.

Consistently with the estimated dietary exposures, BDE-47 is also the predominant PBDE congener in terms of plasma concentrations with a median concentration of 0.588 ng/g of lipids and a large variation in levels across women (min-max: 0.17 to 10.984 ng/g of lipids). The plasma concentrations of BDE-209 were not measured. Another PBDE that we observe in high concentrations in plasma is BDE-153 for which, on the contrary, the estimated dietary exposure is relatively low (median 0.130 ng/kg BW/day, min-max: 0.004-0.032 ng/kg BW/day). For the only polybrominated biphenyl studied (PBB-153), we find relatively high concentrations with a median level of 0.318 ng/g of lipids (min-max: 0.115 and 10.936 ng/g of lipids).

Table IV.3. **Distribution of BFRs concentrations in plasma (ng/g of lipids) and estimated dietary exposure to BFRs (ng/kg BW/day) in our study population (N=168 and N=162 respectively)**

BFRs compounds	Circulating levels				Estimated dietary exposures			
	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.
HBCDalpha					0.054	0.177	0.183	0.499
HBCDbeta					0.004	0.012	0.012	0.027
HBCDgamma					0.009	0.027	0.028	0.055
BDE-28	0.006	0.039	0.057	0.567	0.001	0.006	0.007	0.030
BDE-47	0.170	0.588	0.843	10.984	0.038	0.112	0.125	0.445
BDE-99	0.036	0.133	0.201	4.116	0.017	0.048	0.049	0.109
BDE-100	0.043	0.174	0.247	2.844	0.007	0.021	0.025	0.099
BDE-153	0.219	0.535	0.582	2.317	0.004	0.013	0.014	0.032
BDE-154	0.006	0.029	0.038	0.282	0.004	0.013	0.015	0.054
BDE-183					0.005	0.020	0.021	0.049
BDE-209					0.100	0.311	0.340	0.823
PBDEs					0.195	0.579	0.597	1.395
HBCDs					0.079	0.216	0.223	0.572
PBB-153	0.115	0.318	0.431	10.936				

For hexabromocyclododecanes (HBCDs), dietary exposure was estimated for three congeners with a predominant exposure to the "alpha" congener (median 0.177 ng/kg BW/day, min-max: 0.054-0.499 ng/kg BW/day). Circulating levels of HBCDs were not measured.

When we compared the different congeners to evaluate their correlation (Figure IV.3) we found that for plasma concentrations most correlations are generally weak or moderate (between 0.3 and 0.8) with the exception of plasma concentrations of BDE-47 that are strongly correlated with BDE-99, BDE-100 and BDE-154 (correlations ≥ 0.85) and for BDE-100 that is strongly correlated with BDE-154 (correlation = 0.85). The correlation between congeners is generally weak or moderate also for the estimates of dietary exposure (between 0.2 and 0.8) with some exceptions, notably between BDE-28, BDE-47, BDE-100, and BDE-154 that are virtually perfectly correlated (correlations ≥ 0.99). Interestingly enough, the latter very strong correlation is observed for their plasma concentrations only between BDE-47 and BDE-100 (correlation = 0.93) while between the other congeners correlations of their plasma concentrations are weaker (between 0.5 and 0.85).

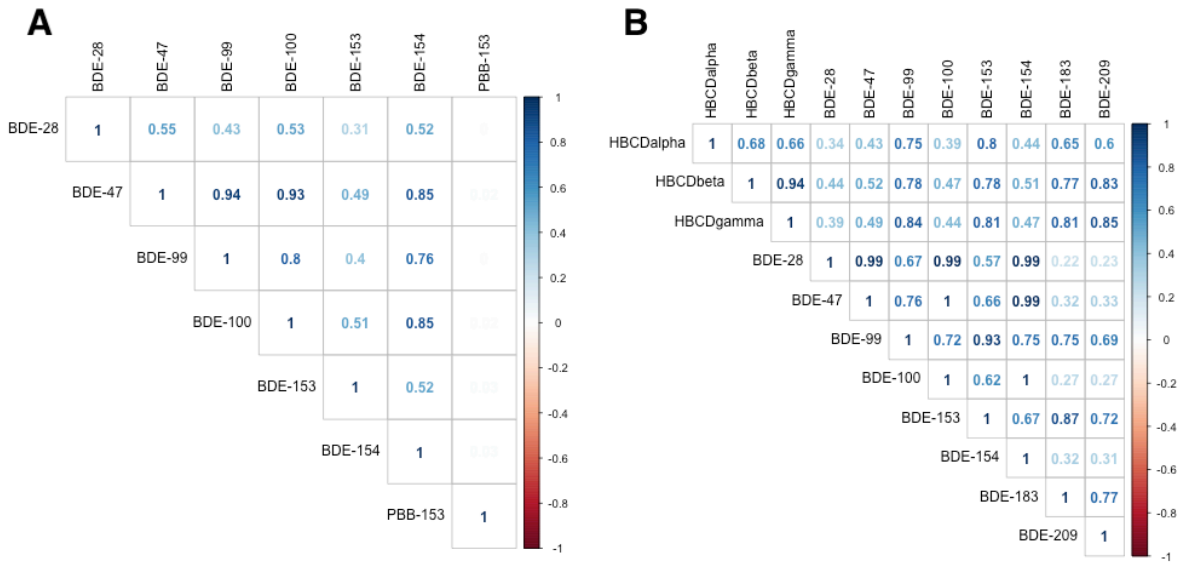


Figure IV.3. **Correlation between the different BFRs congeners for blood concentrations** A) and estimated dietary exposure B) separately.

Even more interestingly, when we compared dietary exposure estimates and measured circulating levels for the 6 PBDEs congeners for which both were available, we found that correlations between the two are very weak and not statistically significant (Table IV.4).

Table IV.4. **Correlations between dietary exposure estimates and circulating levels of PBDEs congeners (N=162)**

Circulating levels	Dietary exposure	Pearson 's correlation	
		Estimates	p-value
BDE-28	BDE-28	0.063	0.421
BDE-47	BDE-47	0.117	0.137
BDE-99	BDE-99	0.087	0.267
BDE-100	BDE-100	0.140	0.073
BDE-153	BDE-153	0.147	0.061
BDE-154	BDE-154	0.107	0.173

3.2.2 EPIGENOME-WIDE ASSOCIATION STUDY: BFRS AND METHYLATION OF BLOOD DNA

For the analyses of the association between BFRs and methylation levels of DNA from blood, we first estimated the association for each individual CpGs (N = 805 837) separately for the estimated dietary exposure to each BFR and for plasma concentrations of each BFR. To take into account the impact of

multiple tests on the level of statistical significance, we assigned such level using the False Discovery Rate (FDR) approach (FDR q-value < 5%)

The quantile-quantile plots with the observed p-values plotted against the expected p-values under the null hypothesis of no association show no evidence of association with circulating levels of BFRs (Figure IV.4) or dietary exposure to BFRs (Figure IV.5) for any of the CpGs with a tendency, for some of the congeners, towards deflation (higher, closer to one, observed p-values relative to expected p-values under the null hypothesis of no association).

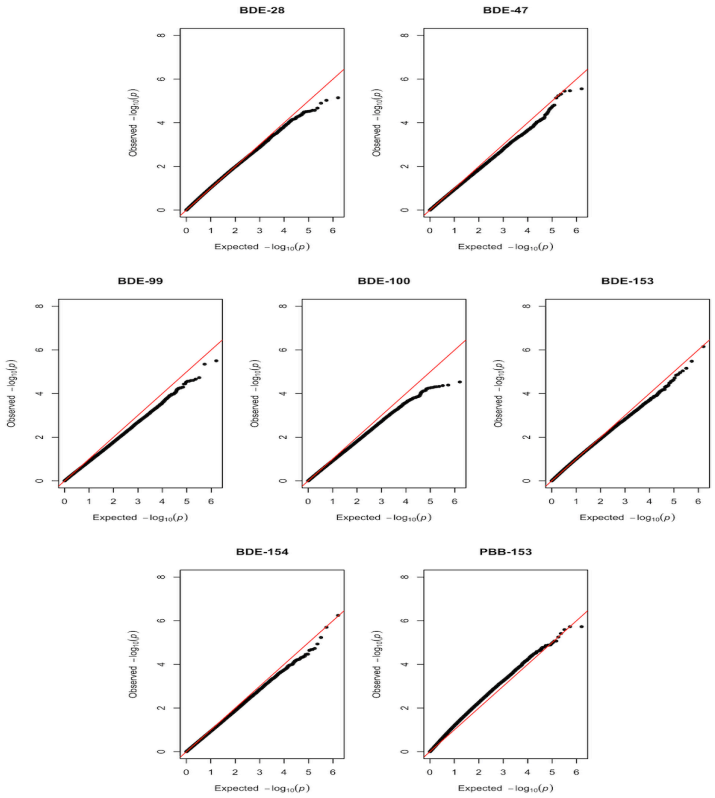


Figure IV.4. **Quantile-quantile plot for the association between circulating levels of BFRs and DNA methylation at 805 837 CpGs sites (N=168)**

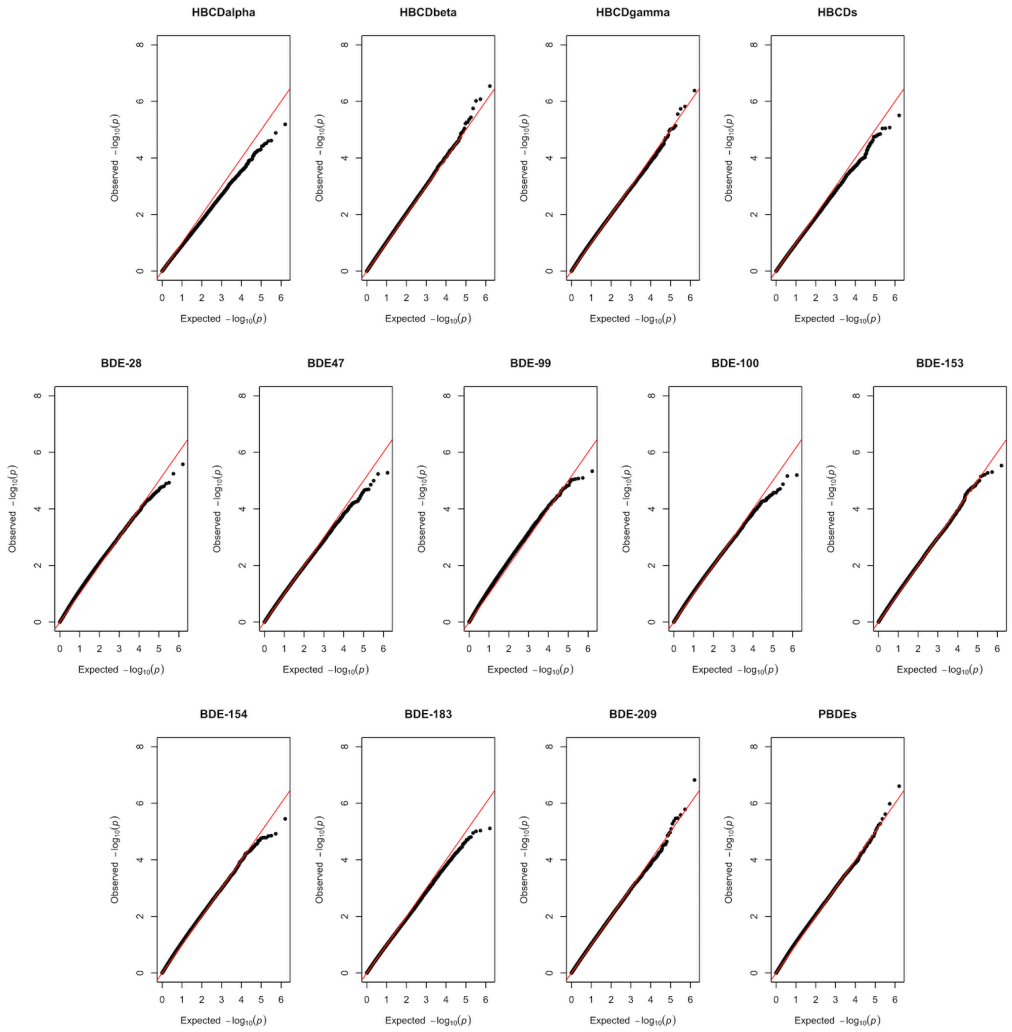


Figure IV.5. **Quantile-quantile plot for association between estimated dietary exposure to BFRs and DNA methylation at 805.837 CpGs sites (N=162)**

For each congener, the top 10 CpG sites (i.e. selected on the basis of the smallest *p*-values) are shown in Appendices 5-8 Interestingly, there is quite a clear tendency in the direction of associations that is distinctly different for dietary exposure to BFRs and plasma concentrations. Most of the regression coefficients are positive for the estimated dietary exposure to BFRs (i.e. higher exposure levels would be associated with higher methylation levels) while they are negative for plasma concentrations (i.e. higher levels would be associated with lower methylation levels).

Despite this interesting tendency, the estimated associations are weak, and none passes the threshold of genome-wide statistical significance. For plasma concentrations, the top CpGs are cg23619365 (β = -0.4, $P = 5.7 \times 10^{-7}$); cg10270519 (β =1.7, $P = 1.8 \times 10^{-6}$) and cg26264999 (β = 0.3, $P = 7.0 \times 10^{-7}$) for BDE-154, PBB-153 and BDE-153 respectively.

For the estimated dietary exposures, the top CpGs are cg06409164 (β =0.2, $P = 1.5 \times 10^{-7}$); cg15267844(β =1.3, $P = 2.4 \times 10^{-7}$) and cg06409164 (β =0.2, $P = 4.4 \times 10^{-7}$) for BDE-209, total PBDEs and HBCDgamma respectively.

Notably, when we compared the top CpGs across the different congeners we found that cg06409164, a CpG located in the body of the gene PARK7 (Parkinsonism associated deglycase) known to be involved in Parkinson's disease, that show a positive association with BDE-209 and HBCDgamma, additionally show a positive association with HBCDbeta (β =1.3, $P = 2.4 \times 10^{-7}$) and PBDEs (β =0.1, $P = 2.4 \times 10^{-5}$).

3.2.3 BFRS AND GLOBAL OR REGIONAL METHYLATION

On the basis of the tendencies observed in the directions of the weak associations for the individual CpGs we calculated an indicator of global DNA methylation equal to the medians in the *M*-values across all CpGs. The distribution of such indicators of global methylation showed a median value of 0.63 ± 0.005. Plasma concentrations were inversely associated with global methylation for all BFRs except BDE-99 but they were statistically significant only for BDE-153 (coefficient β = -0.009, p-value = 4×10^{-2}). In contrast to the results for plasma concentrations, the associations between estimated dietary exposures and global methylation were positive for all congeners with statistically significant associations for HBCDbeta (β = 0.008, p=2.2×10^{-2}), BDE-209 (β = 0.007, p=3.9×10^{-2}) and PBDEs (β = 0.007, p=4×10^{-2}) (Table IV.5).

Table IV.5. **Linear mixed effect models for circulating levels or dietary exposure to BFRs and genome-wide methylation M-value of 805 837 CpGs**

	Circulating levels			Dietary exposure		
	Coefficients[a]	*CI*	*p*	*Coefficients*[b]	*CI*	*p*
HBCDalpha				0.005	-0.003 – 0.012	0.227
HBCDbeta				0.008	0.001 – 0.005	**0.022**
HBCDgamma				0.006	-0.001 – 0.012	0.099
BDE-28	-0.008	-0.018 – 0.003	0.165	0.004	-0.003 – 0.011	0.265
BDE-47	-0.008	-0.021 – 0.006	0.261	0.003	-0.004 – 0.010	0.381
BDE-99	0.002	-0.018 – 0.022	0.826	0.006	-0.001 – 0.012	0.108
BDE-100	-0.010	-0.025 – 0.005	0.182	0.005	-0.003 – 0.012	0.216
BDE-153	-0.009	-0.017 – -0.000	**0.042**	0.005	-0.002 – 0.012	0.184
BDE-154	-0.001	-0.010 – 0.008	0.830	0.003	-0.004 – 0.010	0.406
BDE-183				0.001	-0.006 – 0.008	0.722
BDE-209				0.007	0.000 – 0.013	**0.039**
PBDEs				0.007	0.000 – 0.014	**0.040**
HBCDs				0.005	-0.003 – 0.012	0.228
PBB-153	-0.026	-0.010 – 0.008	0.128			

[a]Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI and parity/total breastfeeding duration as fixed effects
[b]Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern as fixed effects

To explore whether BFRs are associated with altered methylation levels in specific genomic locations selected for relevant functional or spatial characteristics, we used the manifest file provided by Illumina to classify CpGs according to their position relative to CpGs islands (Island/Shore or Shelf/Other), regulatory features (Promoter or Other) and transcription start sites, TSS (TSS1500: within 1500 bps of a transcription start site or TSS200: within 200 bps of a transcription start site).

Overall, consistently with the results for global methylation also the analyses by genomic regions show mostly negative associations for plasma concentrations and positive associations for estimated dietary exposures to BFRs (Table IV.6 and IV.7 for plasma concentrations and Table IV.8 and IV.9 for the estimated dietary exposures). All the estimated associations are at most weak and mostly non-significantly different from the null hypothesis of no association. The strongest evidence of association is between dietary exposure to BDE-209 and methylation levels in promoter regions or shelf regions within promoters (coefficient $\beta = 0.012$, p-value $= 3 \times 10^{-3}$) and between dietary exposure to PBDEs and methylation levels in CpG islands and shores (coefficient $\beta = 0.010$, p-value $= 5 \times 10^{-3}$).

Table IV.6. **Linear mixed effect models for circulating levels of BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions**

	Island or Shore			Shelf or None			Promoter			Other		
	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*
BDE-28	-0.003	-0.014 – 0.009	0.666	-0.013	-0.030 – 0.005	0.158	-0.011	-0.024 – 0.002	0.086	-0.006	-0.017 – 0.005	0.285
BDE-47	-0.007	-0.021 – 0.008	0.378	-0.011	-0.033 – 0.011	0.327	-0.008	-0.024 – 0.009	0.365	-0.008	-0.022 – 0.006	0.273
BDE-99	-0.006	-0.027 – 0.016	0.604	0.009	-0.024 – 0.041	0.600	-0.002	-0.026 – 0.023	0.902	-0.001	-0.022 – 0.020	0.917
BDE-100	-0.010	-0.026 – 0.006	0.211	-0.013	-0.038 – 0.011	0.288	-0.012	-0.030 – 0.007	0.211	-0.011	-0.027 – 0.004	0.151
BDE-153	-0.011	-0.020 – -0.002	**0.015**	-0.011	-0.025 – 0.003	0.134	-0.010	-0.020 – -0.000	**0.044**	-0.011	-0.020 – -0.002	**0.015**
BDE-154	-0.004	-0.014 – 0.006	0.410	0.001	-0.014 – 0.016	0.932	-0.009	-0.020 – 0.002	0.109	-0.002	-0.012 – 0.008	0.668
PBB-153	-0.043	-0.078 – -0.007	**0.019**	-0.017	-0.072 – 0.038	0.541	-0.033	-0.073 – 0.008	0.115	-0.034	-0.069 – 0.001	0.059

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI and parity/total breastfeeding duration as fixed effects

Table IV.7. **Linear mixed effect models for dietary exposure to BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions**

	Island or Shore			Shelf or None			Promoter			Other		
	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*
HBCDalpha	0.001	-0.007 – 0.009	0.731	0.009	-0.003 – 0.021	0.149	0.005	-0.004 – 0.014	0.289	0.005	-0.003 – 0.012	0.254
HBCDbeta	0.007	-0.001 – 0.014	0.078	0.012	0.001 – 0.023	**0.034**	0.008	-0.000 – 0.016	0.063	0.008	0.001 – 0.015	**0.026**
HBCDgamma	0.008	0.001 – 0.015	**0.026**	0.007	-0.004 – 0.018	0.192	0.010	0.002 – 0.018	**0.015**	0.007	-0.000 – 0.014	0.052
BDE-28	0.008	0.001 – 0.016	**0.025**	0.003	-0.009 – 0.015	0.645	0.008	-0.001 – 0.016	0.069	0.006	-0.001 – 0.014	0.111
BDE-47	0.007	-0.000 – 0.014	0.057	0.002	-0.010 – 0.013	0.753	0.007	-0.001 – 0.015	0.082	0.005	-0.002 – 0.012	0.185
BDE-99	0.006	-0.001 – 0.014	0.078	0.008	-0.003 – 0.019	0.161	0.008	-0.000 – 0.016	0.057	0.006	-0.001 – 0.014	0.074
BDE-100	0.008	0.000 – 0.015	**0.045**	0.004	-0.007 – 0.016	0.470	0.007	-0.001 – 0.015	0.104	0.006	-0.001 – 0.014	0.106
BDE-153	0.008	0.000 – 0.016	**0.045**	0.007	-0.005 – 0.019	0.272	0.010	0.001 – 0.019	**0.027**	0.007	-0.001 – 0.014	0.093
BDE-154	0.008	0.000 – 0.015	**0.049**	0.002	-0.010 – 0.014	0.786	0.007	-0.002 – 0.015	0.109	0.005	-0.002 – 0.013	0.176
BDE-183	0.003	-0.004 – 0.010	0.416	0.002	-0.009 – 0.014	0.658	0.005	-0.003 – 0.013	0.248	0.001	-0.006 – 0.008	0.712
BDE-209	0.009	0.002 – 0.015	**0.013**	0.009	-0.001 – 0.020	0.078	0.012	0.004 – 0.019	**0.003**	0.008	0.001 – 0.015	**0.023**
PBDEs	0.010	0.003 – 0.018	**0.005**	0.008	-0.003 – 0.020	0.138	0.010	0.002 – 0.019	**0.013**	0.010	0.002 – 0.017	**0.009**
HBCDs	0.001	-0.007 – 0.009	0.836	0.010	-0.003 – 0.023	0.119	0.005	-0.005 – 0.014	0.325	0.004	-0.004 – 0.012	0.300

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern as fixed effects

Table IV.8. **Linear mixed effect models for circulating levels of BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	TSS1500 or TSS200						Promoter					
	Island or Shore			Shelf or None			Island and Shore			Shelf or None		
	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*
BDE-28	-0.001	-0.014 – 0.012	0.860	-0.012	-0.027 – 0.003	0.112	-0.006	-0.023 – 0.011	0.504	-0.015	-0.028 – -0.002	**0.026**
BDE-47	-0.006	-0.023 – 0.011	0.483	-0.012	-0.032 – 0.007	0.200	-0.009	-0.031 – 0.013	0.420	-0.007	-0.025 – 0.010	0.409
BDE-99	-0.006	-0.031 – 0.019	0.622	0.003	-0.025 – 0.032	0.815	-0.006	-0.039 – 0.026	0.705	0.001	-0.025 – 0.026	0.957
BDE-100	-0.009	-0.028 – 0.009	0.332	-0.015	-0.036 – 0.006	0.159	-0.012	-0.036 – 0.012	0.322	-0.012	-0.031 – 0.007	0.215
BDE-153	-0.011	-0.022 – -0.001	**0.032**	-0.012	-0.024 – 0.000	0.054	-0.012	-0.026 – 0.001	0.072	-0.010	-0.020 – 0.000	0.061
BDE-154	-0.004	-0.015 – 0.007	0.452	-0.002	-0.015 – 0.011	0.749	-0.009	-0.024 – 0.005	0.213	-0.009	-0.021 – 0.002	0.106
PBB-153	-0.047	-0.088 – -0.006	**0.023**	-0.021	-0.068 – 0.027	0.392	-0.052	-0.106 – 0.001	0.055	-0.024	-0.066 – 0.018	0.262

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI and parity/total breastfeeding duration as fixed effects

Table IV.9. **Linear mixed effect models for dietary exposure to BFRs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	TSS1500 or TSS200						Promoter					
	Island or Shore			Shelf or None			Island and Shore			Shelf or None		
	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*	*Coefficients**	*CI*	*p*
HBCDalpha	0.000	-0.009 – 0.009	0.977	0.007	-0.003 – 0.018	0.183	0.002	-0.010 – 0.013	0.778	0.007	-0.002 – 0.016	0.138
HBCDbeta	0.007	-0.001 – 0.015	0.106	0.011	0.002 – 0.021	**0.022**	0.008	-0.003 – 0.019	0.133	0.008	-0.001 – 0.017	0.073
HBCDgamma	0.009	0.001 – 0.017	**0.023**	0.008	-0.002 – 0.017	0.106	0.012	0.002 – 0.022	**0.023**	0.009	0.001 – 0.017	**0.032**
BDE-28	0.010	0.002 – 0.018	**0.016**	0.004	-0.006 – 0.014	0.447	0.013	0.003 – 0.024	**0.015**	0.005	-0.004 – 0.013	0.274
BDE-47	0.009	0.000 – 0.017	**0.041**	0.003	-0.007 – 0.013	0.575	0.012	0.001 – 0.022	**0.032**	0.005	-0.003 – 0.014	0.232
BDE-99	0.007	-0.001 – 0.015	0.083	0.009	-0.001 – 0.018	0.076	0.010	-0.000 – 0.021	0.056	0.007	-0.001 – 0.016	0.103
BDE-100	0.009	0.001 – 0.017	**0.035**	0.005	-0.005 – 0.015	0.341	0.012	0.001 – 0.022	**0.034**	0.004	-0.004 – 0.013	0.30
BDE-153	0.009	0.000 – 0.018	**0.040**	0.007	-0.003 – 0.018	0.163	0.011	-0.000 – 0.023	0.055	0.010	0.001 – 0.019	**0.038**
BDE-154	0.00[2–4]9	0.001 – 0.018	**0.033**	0.003	-0.007 – 0.013	0.577	0.012	0.001 – 0.023	**0.034**	0.004	-0.004 – 0.013	0.316
BDE-183	0.005	-0.003 – 0.013	0.245	0.003	-0.006 – 0.013	0.494	0.006	-0.004 – 0.017	0.231	0.004	-0.004 – 0.013	0.340
BDE-209	0.009	0.001 – 0.017	**0.023**	0.010	0.001 – 0.019	**0.025**	0.012	0.002 – 0.022	**0.021**	0.012	0.004 – 0.020	**0.003**
PBDEs	0.011	0.003 – 0.019	**0.007**	0.009	-0.001 – 0.019	0.067	0.013	0.003 – 0.024	**0.015**	0.009	0.001 – 0.018	**0.030**
HBCDs	-0.001	-0.010 – 0.009	0.914	0.008	-0.003 – 0.019	0.145	0.000	-0.012 – 0.013	0.956	0.007	-0.002 – 0.017	0.135

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern as fixed effects

3.2.4 BFRS AND METHYLATION ALTERATION IN SPECIFIC PATHWAYS: GENE SET ENRICHMENT ANALYSES

The gene set enrichment analyses that we performed to identify specific pathways in which gene belonging to such pathway are differentially methylated according to the levels of BFRs, provide evidence for altered methylation in pathways that are distinct for plasma concentrations and estimated dietary exposure (Table IV.10).

For plasma concentrations three of the four gene sets identified are positively enriched: BDE-47 is associated with gene enrichment of "DNA repair" and "IL6-JAK-STAT3 signaling" and BDE-154 with "androgen response". Positive enrichment means that the levels of DNA methylation of the genes included in the gene set are positively correlated with the plasma concentrations of the corresponding BFRs. These results suggest that plasma concentrations of BFRs may be associated with increased methylation levels in genes in pathways involved in signaling in processes such as immune response and cell cycle regulation ("IL6-JAK-STAT3"), androgen response and DNA repair. The negative correlation between plasma concentrations of BDE-28 and methylation levels in genes in the gene set "MYC targets" is of particular interest as MYC is a proto-oncogene.

For dietary exposures to BFRs the three gene sets identified do not overlap with those identified for plasma concentrations: HBCDalpha and total HBCDs are associated with negative enrichment of the gene set "Apoptosis" and HBCDbeta with negative enrichment of "TNFalpha signaling via NK-kB" that includes genes regulated by the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) in response to tumour necrosis factor (TNFalpha), a potent cytokine and critical regulator of apoptosis, inflammation, and immunity via control of the transcription factor NF-κB. On the contrary, BDE-183 is associated with positive enrichment of the gene set "Hypoxia" including genes involved in the response to low levels of oxygen.

Table IV.10. **Gene set enrichment analysis results for genes that are positively or negatively correlated to BFRs exposure**
Only gene sets for which the FDR q-value < 0.3 are provided.

	Circulating levels				**Dietary exposure**			
	Gene set (number of genes identified)	*ES*	*p*	*FDR*	*Gene set (number of genes identified)*	*ES*	*p*	*FDR*
HBCDalpha					APOPTOSIS (25)	-0.301	0.007	0.139
HBCDbeta					TNFA_SIGNALING_VIA_NFKB (46)	-0.233	0.011	0.179
HBCDgamma								
BDE-28	MYC_TARGETS_V1 (57)	-0.212	0.007	0.197				
BDE-47	DNA_REPAIR (28)	0.269	0.021	0.198				
	IL6_JAK_STAT3_SIGNALING (17)	0.362	0.011	0.204				
BDE-99								
BDE-100								
BDE-153								
BDE-154	ANDROGEN_RESPONSE (30)	0.260	0.025	0.263				
BDE-183					HYPOXIA (24)	0.271	0.047	0.290
BDE-209								
PBDEs								
HBCDs					APOPTOSIS (33)	-0.269	0.013	0.251
PBB-153								

4. METHYLATION SIGNATURES OF PER- AND POLYFLUORINATED ALKYLATED SUBSTANCES

4.1 APPROACHES

4.1.1 ASSOCIATION BETWEEN DIETARY EXPOSURE TO PFASS AND DNA METHYLATION

As conducted for BFRs, women with aberrant energy intake (e.g. 1% and 99% extremes of the energy intake/energy expenditure ratio) were excluded (n=6) and our final dataset for the association between DNA methylation and dietary exposure to PFASs consisted of a subset of 162 women with methylation data on 805.837 CpGs.

Then, we explored the association between DNA methylation and dietary exposure to PFASs (PFOA and PFOS) through several linear mixed-effects models with DNA methylation as dependent variable (either global methylation, or “regional” methylation or single probes), quartiles of PFASs as explanatory variable with plate and chips considered as random effects. Additionally, to what have been done for BFRs, models were fitted with adjustment for lipids (categorical, below or above the median).

Models were adjusted for dietary patterns as they are potential confounders because they are associated with both "exposure" to PFASs (or rather the proxy used in our analyzes, which is calculated precisely from diet), and potentially methylation. In the same logic, we decided to adjust for lipids for which two approaches are generally used in the literature; those using measurements in "ng/g of lipids" and in "ng/ml of serum/plasma". If the majority of the authors agree on the use of "ng/g of lipids" for BFRs, in particular because of their lipophilic characteristics, the proposals are rather divergent compared to PFASs. Rather, these substances tend to accumulate in tissues such as the liver, and some studies suggest a disruption of the lipid regulatory mechanisms[190,191].

These adjustments were discussed and defined with Francesca Mancini, the team's coordinator of research on food contaminants, in accordance with the literature and the approaches used for previous studies / explorations on the same exposure data for BFRs. and to PFASs.

4.1.2 ASSOCIATION BETWEEN CIRCULATING LEVELS OF PFASS AND DNA METHYLATION

For the analyses of PFASs (PFOA and PFOS) blood levels, that we conducted separately to the analyses of dietary exposure to PFASs, data were available for a slightly larger sample of women (N=166). For such analyses we used models similar to those used for the analyses of dietary exposure with the exception of the adherence scores to the healthy dietary pattern and the Western dietary pattern that were not included in the models adjusted for the covariates.

4.1.3 ENRICHMENT ANALYSIS

To determine whether any gene set or biological pathway is overrepresented in the list of genes whose DNA methylation are associated with circulating levels or dietary exposure to PFASs, we performed GSEA using an approach similar to the one used in the analyses conducted for BFRs: (1) genes near CpG sites located in promoter region in which the association between circulating levels of PFASs and CpG site methylation levels are significant and below 5% and (2) genes near CpG sites located in promoter region in which the association between dietary exposure to PFASs and CpG site methylation levels are significant and below 5%.

4.2 FINDINGS

4.2.1 BASELINE CHARACTERISTICS OF THE STUDY POPULATION

The baseline characteristics of study participants are summarized in Table IV.11. To study the association between PFASs and methylation of DNA from blood, data were available from 166 women for circulating levels of PFASs and from 162 women for the dietary exposure to these compounds. Median age of the study participants was 56.1 years and most of them had a healthy body mass index with only one quarter of them being overweight or obese. About 43% of them are nulliparous or never breastfed, 40% had at least one child but breastfed for less than 6 months and 16.5% had breastfed for more than 6 months.

From the detailed data from the food frequency questionnaire completed in 1993, between 2 and 5 years before the blood collection, dietary patterns were identified including a “healthy” dietary pattern and a “Western” dietary pattern[189]. In our study population around half of the women had a “healthy” diet and half adhered to a Western diet with a small overlap between the two groups.

Table IV.11. **Baseline characteristics of the study population**

	Circulating levels (n = 166)	Dietary exposure (n = 162)
Age (%)		
<56.1	81 (48.8)	79 (48.8)
>56.1	85 (51.2)	83 (51.2)
Body Mass Index (%)		
<25	122 (73.5)	121 (74.7)
>25	44 (26.5)	41 (25.3)
Score of adherence to the healthy dietary pattern (%)		
Above median		87 (53.7)
Below median		75 (46.3)
Score of adherence to the Western dietary pattern (%)		
Above median		82 (50.6)
Below median		80 (49.4)
Parity and total breastfeeding duration (%)		
Nulliparous or never breastfeed	73 (44.0)	70 (43.2)
Parous and breastfeed for less than 6 months	66 (39.8)	65 (40.1)
Parous and breastfeed for more than 6 months	27 (16.3)	27 (16.7)
Lipids		
Above median	86 (51.8)	84 (51.9)
Below median	80 (48.2)	78 (48.1)

The levels of dietary exposure to PFASs estimated in our study population are presented in Table IV.12. For circulating levels of PFOA, the median concentration is 6.83 ng/mL (min-max: 1.287 to 17.685 ng/L), while for PFOS the median of concentration is 17.32 ng/mL (min-max: 6.612 to 59.119 ng/mL

The median dietary exposure to PFOS and to PFOA was respectively 0.443 ng/kg BW/day (min-max: 0.108 to 1.441 ng/kg BW/day) and 0.132 ng/kg BW/day (min-max: 0.132 to 1.342 ng/kg BW/day) respectively.

Table IV.12. **Distribution of PFASs concentrations in serum (ng/mL) and estimated dietary exposure to PFASs (ng/kg BW/day) in our study population (N=168 and N=162 respectively)**

PFASs compounds	**Circulating levels**				**Dietary exposures**			
	Min.	Median	Mean	Max.	Min.	Median	Mean	Max.
PFOA	1.287	6.831	7.263	17.685	0.108	0.443	0.486	1.441
PFOS	6.612	17.320	18.694	59.119	0.132	0.506	0.530	1.342

Additionally, strong correlations were observed between dietary intakes of PFOA and PFOS (0.94) while a moderate value is observed for the correlation between circulating levels of PFOA and PFOS (0.54) (Table IV.13).

Table IV.13. **Correlation between the different PFASs congeners for blood concentrations and estimated dietary exposure separately**

			Pearson 's correlation	
			Estimates	p
Circulating levels	PFOA	PFOS	0.535	< 0.001
Dietary exposure	PFOA	PFOS	0.84	< 0.001

With the regards to the correlation between these compounds estimated from diet in comparison to circulating levels, inverse and weak correlations are observed (Table IV.14) with regard to PFOA ($p = 1.2 \times 10^{-2}2$) and PFOS ($p = 5.78\times 10^{-2}$).

Table IV.14**. Correlation between dietary exposure estimates and circulating levels of PFASs congeners (N = 162)**

Circulating levels	Dietary exposure	Pearson 's correlation	
		Estimates	p
PFOA	PFOA	-0.198	0.012
PFOS	PFOS	-0.044	0.578

4.2.2 EPIGENOME-WIDE ASSOCIATION STUDY: PFASS AND METHYLATION OF BLOOD DNA

For the analyses of the association between PFASs and methylation levels of DNA from blood, we first estimated the association for each individual CpGs (N = 805 837) separately for the estimated dietary exposure to each PFAS and for plasma concentrations of each PFAS. To take into account the impact of multiple tests on the level of statistical significance, we assigned such level using the False Discovery Rate (FDR) approach (FDR q-value < 5%).

The quantile-quantile plots with the observed p-values plotted against the expected p-values under the null hypothesis of no association show no evidence of association with circulating levels of PFASs (Figure IV.6 A)) or dietary exposure to PFASs (Figure IV.6 B)) for any of the CpGs with a tendency, for some of the congeners, towards deflation (higher, closer to one, observed p-values relative to expected p-values under the null hypothesis of no association).

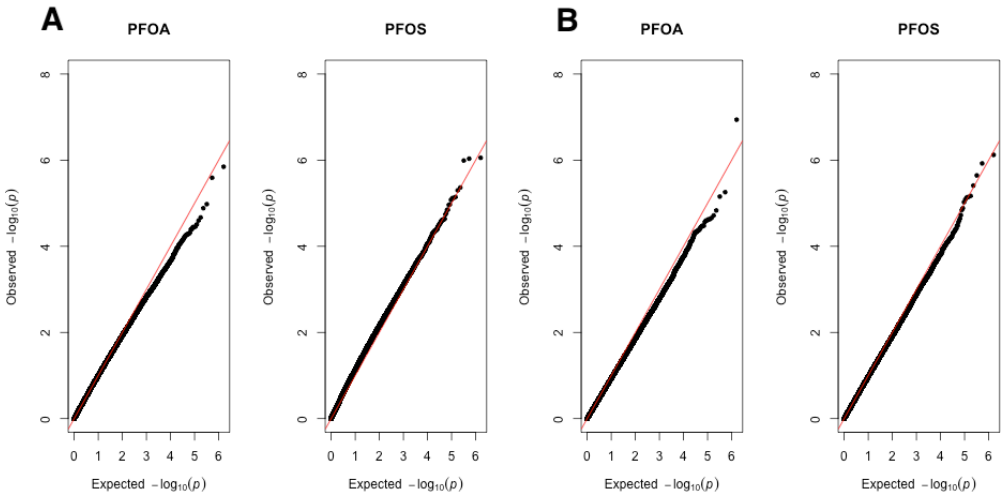


Figure IV.6. **Quantile-quantile plot for association between circulating levels of PFASs and dietary exposure to PFASand DNA methylation at 805.837 CpGs sites**
A) circulating levels of PFASs. B) dietary exposure to PFASand DNA methylation at 805.837 CpGs sites
For each congener, the top 10 CpG sites (i.e. selected on the basis of the smallest *p*-values) are shown in Appendices 9 and 10.

Interestingly, there is quite a clear tendency in the direction of associations that is distinctly different for dietary exposure to PFASs and plasma concentrations. Most of the regression coefficients are positive for the estimated dietary exposure to PFASs (i.e. higher exposure levels would be associated with higher methylation levels) while they are negative for plasma concentrations (i.e. higher levels would be associated with lower methylation levels).

Despite this interesting tendency, the estimated associations are weak, and none passes the threshold of genome-wide statistical significance. For plasma concentrations, the top CpGs are, cg06874740 (β = -0.37, $p = 1.42 \times 10^{-6}$) and cg15913831 (β = -0.401, $p = 8.8 \times 10^{-7}$) for PFOA and PFOS respectively. For the estimated dietary exposures, the top CpGs are cg08255137 (β = 0.2, $p = 1.49 \times 10^{-7}$) and cg25246012 (β = 0.255, $p = 7.5 \times 10^{-7}$) for PFOA and PFOS respectively.

4.2.3 PFASS AND GLOBAL OR REGIONAL METHYLATION

On the basis of the tendencies observed in the directions of the weak associations for the individual CpGs we calculated an indicator of global DNA methylation equal to the medians in the *M*-values across all CpGs. The distribution of such indicators of global methylation showed a median value of 0.63 ± 0.005. Plasma concentrations were inversely associated with global methylation for PFOA (β = -0.003, $p = 3.26 \times 10^{-1}$) and PFOS (β = -0.001, $p = 7.18 \times 10^{-1}$) (Table IV.15).

In contrast to the results for plasma concentrations, the associations between estimated dietary exposures and global methylation were positive for both PFOA (β = 0.001, p = 7.73×10^{-1}) and PFOS (β = 0.002, p = 5.87×10^{-1}).

Table IV.15. **Linear model for circulating levels or dietary exposure to PFASs and genome-wide methylation of 805.837 CpGs**

	Circulating levels[a]			**Dietary exposure[b]**		
	Coefficients[a]	*CI*	*p*	*Coefficients[b]*	*CI*	*p*
PFOA	-0.003	-0.010 – 0.003	0.326	0.001	-0.005 – 0.007	0.773
PFOS	-0.001	-0.008 – 0.006	0.718	0.002	-0.005 – 0.009	0.587

[a]Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration and lipids as fixed effects
[b]Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern and lipids as fixed effects

Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern and lipids as fixed effects.

To explore whether PFASs are associated with altered methylation levels in specific genomic locations selected for relevant functional or spatial characteristics, we used the manifest file provided by Illumina to classify CpGs according to their position relative to CpGs islands (Island/Shore or Shelf/Other), regulatory features (Promoter or Other) and transcription start sites, TSS (TSS1500: within 1500 bps of a transcription start site or TSS200: within 200 bps of a transcription start site).

Overall, consistently with the results for global methylation also the analyses by genomic regions show mostly negative associations for plasma concentrations and positive associations for estimated dietary exposures to BFRs (Table IV.16 and IV.17 for plasma concentrations and Table IV.18 and IV.19 for the estimated dietary exposures). All the estimated associations are at most weak and mostly non-significantly different from the null hypothesis of no association.

Overall, we observed inverse and non-significant associations between circulating levels of PFASs and methylation at CpGs in all genomic regions except those located in or near a CGI without difference for PFOA or PFOS respectively in regard to TSS1500 or TSS200 (β = 0.002, p=5.77×10^{-1}; β = 0.005, p=2.93×10^{-1}) or Promoter region (β = 0.004, p=4.81×10^{-1}; β = 0.005, p=3.75×10^{-2}).We observed positive association between PFOS and DNA methylation in or near the CGI (β = 0.008, p=3.8×10^{-2}) which remain significant (Table IV.19) within TSS1500 or TSS200 (β = 0.009, p=4.7×10^{-2}). Additionally, we also observe positive association between Promoter (β = 0.009, p=4.8×10^{-2}) and Shelf or others region (β = 0.02, p=4.9×10^{-2}).

Table IV.16. **Linear mixed effect models for circulating levels of PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	Island or Shore			Shelf or None			Promoter			Other		
	Coefficients	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*
PFOA	0.001	-0.006 – 0.008	0.816	-0.007	-0.017 – 0.004	0.211	-0.000	-0.008 – 0.008	0.985	-0.003	-0.010 – 0.004	0.426
PFOS	0.003	-0.005 – 0.010	0.477	-0.005	-0.016 – 0.007	0.430	-0.000	-0.009 – 0.008	0.915	0.000	-0.007 – 0.008	0.993

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration and lipids as fixed effects

Table IV.17. **Linear mixed effect models for dietary exposure to PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	TSS1500 or TSS200						Promoter					
	Island or Shore			Shelf or None			Island and Shore			Shelf or None		
	Coefficients	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*
PFOA	0.002	-0.006 – 0.010	0.577	-0.005	-0.015 – 0.004	0.243	0.004	-0.007 – 0.015	0.481	-0.002	-0.011 – 0.006	0.557
PFOS	0.005	-0.004 – 0.013	0.293	-0.004	-0.014 – 0.006	0.424	0.005	-0.006 – 0.017	0.375	-0.004	-0.013 – 0.005	0.435

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern and lipids as fixed effects

Table IV.18. **Linear mixed effect models for circulating levels of PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	Island or Shore			Shelf or None			Promoter			Other		
	Coefficients	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*
PFOA	0.002	-0.005 – 0.008	0.604	0.003	-0.014 – 0.020	0.738	0.003	-0.004 – 0.011	0.399	0.002	-0.005 – 0.008	0.608
PFOS	0.008	0.000 – 0.016	**0.038**	0.020	0.000 – 0.040	**0.049**	0.009	0.000 – 0.018	**0.048**	0.005	-0.003 – 0.013	0.211

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration and lipids as fixed effects

Table IV.19. **Linear mixed effect models for dietary exposure to PFASs and median M-values across regions defined on the basis of their position relative to CpG islands and across functional genomic regions.**

	TSS1500 or TSS200						Promoter					
	Island or Shore			Shelf or None			Island and Shore			Shelf or None		
	Coefficients	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*	*Coefficients*	*CI*	*p*
PFOA	0.001	-0.007 – 0.009	0.773	0.002	-0.006 – 0.011	0.585	0.001	-0.009 – 0.011	0.803	0.004	-0.003 – 0.012	0.261
PFOS	0.009	0.000 – 0.018	**0.047**	0.002	-0.008 – 0.013	0.679	0.011	-0.000 – 0.023	0.057	0.008	-0.001 – 0.017	0.084

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern and lipids as fixed effects

4.2.4 PFASS AND METHYLATION ALTERATIONS IN SPECIFIC PATHWAYS: GENE SET ENRICHMENT ANALYSIS

The gene set enrichment analyses that we performed to identify specific pathways in which gene belonging to such pathway are differentially methylated according to the levels of PFASs, provide evidence for altered methylation in pathways that are distinct for plasma concentrations and estimated dietary exposure (Table IV.20).

For plasma concentrations five of the six gene sets identified are positively enriched: PFOA is associated with gene enrichment of "Myc_targets_v2" and "Hypoxia" while PFOS with "Il2_Stat5 signaling", "cholesterol homeostasis", "inflammatory response" and "fatty acid metabolism".

Positive enrichment means that the levels of DNA methylation of the genes included in the gene set are positively correlated with the plasma concentrations of the corresponding PFASs. These results suggest that plasma concentrations of PFASs may be associated with increased methylation levels in genes in pathways involved in processes such as immune and inflammatory response; cholesterol and fatty acid. The negative correlation between plasma concentrations of PFOA and methylation levels in genes in the gene set "Hypoxia" is of particular interest as it represents a set of genes up-regulated in response to low oxygen levels.

For dietary exposures to PFASs, particularly PFOA, the only gene set identified do not overlap with those identified for plasma concentrations which is associated with negative enrichment of the gene set "Apoptosis".

Table IV.20. **Gene set enrichment analysis results for genes that are positively or negatively correlated to PFASs exposure (FDR < 0.3)**
Only gene sets for which the FDR q-value < 0.3 are provided.

	Circulating levels				Dietary exposure			
	Gene set	*ES*	*P*	*FDR*	*Gene set*	*ES*	*P*	*FDR*
PFOA	MYC_TARGETS_V2 (22)	0.333	*0.016*	0.266	APOPTOSIS (28)	-0.313	0.003	0.063
	HYPOXIA (49)	-0.206	0.019	0.269				
PFOS	IL2_STAT5_SIGNALING (56)	0.191	0.030	0.184				
	CHOLESTEROL_HOMEOSTASIS (25)	0.279	0.031	0.198				
	INFLAMMATORY_RESPONSE (33)	0.251	0.035	0.221				
	FATTY_ACID_METABOLISM (37)	0.240	0.033	0.277				

5. CONCLUSION

5.1 METHYLATION SIGNATURES OF BROMINATED FLAME RETARDANTS

BFRs exposure has become an increasingly important global public health given contamination in the environment, their tendency to bioaccumulate in human tissue and their effects on biological systems that are yet to be fully elucidated. Our hypothesis is that, among other impacts on humans, BFRs may alter methylation levels in human DNA and through such alterations BFRs would exert multiple actions on human health. To test some aspects of this hypothesis, we used blood DNA from a sample of 162-168 women from our prospective E3N cohort. Individual CpG analyses and analyses of global and regional DNA methylation did not provide convincing evidence of associations with BFRs plasma concentrations or dietary exposure to BFRs.

The results obtained from the gene enrichment analyses are interesting as they show that exposure to BFRs may alter the levels of circulating DNA methylation in specific pathways. Plasma concentrations and dietary exposure to BFRs appear to be associated with DNA methylation alterations in different pathways. While for BFRs circulating levels the identified gene sets enriched are involved in embryological development, regulation of extracellular matrix, acute phase response, cell cycle regulation and DNA repair mechanisms, for dietary exposure they are related to immune response, hypoxia and apoptosis. These results are somehow broadly consistent with the capacity of BFRs to alter the endocrine system, influence the immune response and impact on the reproductive system in humans and provide support to previous reports that indicate that individual PBDEs and their mixtures can shift cytokine production to a more pro-inflammatory phenotype[192,193] and lead to adverse effects on the reproductive development[194,195].

The different results for BFRs plasma concentrations and the estimated dietary exposures may be explained by the fact that the two estimates of exposure to BFRs are quite distinct. One, obtained from food frequency questionnaires data in 1993 as well as levels of contaminants from the ANSES survey, is an estimate of the exposure through diet, the main source of exposure, while the other is a direct measure of circulating levels in blood samples collected a few years after the questionnaire (1995-1998). It is important to note that circulating levels of BFRs are determined by a complex interplay of factors including exposure from multiple sources (e.g. diet, dust or other environmental sources) but also from the rate of elimination of BFRs through human matrices, in particular, through breastfeeding.

To our knowledge, this is the first epigenome-wide association study of BFRs and DNA methylation. As mentioned earlier (see chapter I) previous studies showed that endocrine disruptors such as phthalates

or bisphenols were associated with hypomethylation, but such studies focused on repetitive genomic elements that were used as markers of global methylation (i.e. Alu and LINE-1). Our study measured DNA methylation in a more systematic manner with coverage of almost 1 million individual CpGs representing more than 90% of all CpGs – i.e. a coverage 6 times greater than the coverage of studies that used Alu and LINE-1 elements.

The main limitations of our study include the cross-sectional nature of the measures in blood (i.e. BFRs plasma concentrations and DNA methylation were measured from the same blood samples) and the relatively limited sample size. Also, we cannot exclude that BFRs influence DNA methylation in other target tissues that were not available for this study.

In conclusion, our study found no evidence of association between BFRs exposure and moderate or strong global or single CpG alterations in circulating DNA methylation. The suggestive evidence of association between BFRs exposure and DNA methylation alterations in specific gene pathways warrant replication in independent studies but it is intriguing as it might reflect a more complex action of this class of substances.

5.2 METHYLATION SIGNATURES OF PER- AND POLYFUORINATED ALKYLATED SUBSTANCES

As BFRs, PFASs exposure is a worldwide concern and we hypothesize that PFAs may alter levels in human DNA and through such alterations PFASs would exert multiple actions on human health. To test some aspects of this hypothesis, we used blood DNA from a sample of 162-166 women from our prospective E3N cohort.

Individual CpG analyses and analyses of global and regional DNA methylation did not provide convincing evidence of associations with PFASs plasma concentrations or dietary exposure to PFASs.

The results obtained from the gene enrichment analyses are interesting as they show that exposure to PFASs may alter the levels of circulating DNA methylation in specific pathways. Plasma concentrations and dietary exposure to PFASs appear to be associated with DNA methylation alterations in different pathways. While for PFASs circulating levels the identified gene sets enriched are involved immune and inflammatory response, cholesterol homeostasis and fatty acid metabolism for dietary exposure they are related to apoptosis.

These results are somehow broadly consistent with the capacity of PFASs, particularly PFOS to activate nuclear receptors such as PPAR-α and induce peroxisome proliferation[91] and influence the immune response or disrupt lipid metabolism and hepatotoxicity[196–198].

Similarly, the different results for PFASs plasma concentrations and the estimated dietary exposures may be explained by the fact that the two estimates of exposure to PFASs are quite distinct. It is also important to note that circulating levels of PFASs are determined by a complex interplay of factors including exposure from multiple sources (e.g. diet, dust or other environmental sources) but also from the rate of elimination of PFASs through human matrices.

To our knowledge, this is the first epigenome-wide association study of PFASs and DNA methylation. As mentioned earlier, a study showed that prenatal exposure to PFOS was associated with hypomethylation, but it was focused on repetitive genomic elements that were used as markers of global methylation (i.e. Alu and LINE-1)[106]. Our study measured DNA methylation in a more systematic manner with coverage of almost 1 million individual CpGs representing more than 90% of all CpGs – i.e. a coverage 6 times greater than the coverage of studies that used Alu and LINE-1 elements.

The main limitations of our study include the cross-sectional nature of the measures in blood (i.e. PFASs plasma concentrations and DNA methylation were measured from the same blood samples) and the relatively limited sample size. Also, we cannot exclude that PFASs influence DNA methylation in other target tissues such as liver that were not available for this study.

In conclusion, our study found no evidence of association between PFASs exposure and moderate or strong global or single CpG alterations in circulating DNA methylation. Additionally, the suggestive evidence of association between PFASs exposure and DNA methylation alterations in specific gene pathways warrant replication in independent studies but it is intriguing as it might reflect a more complex action of this class of substances.

CHAPTER V:

GENERAL DISCUSSION AND FUTURE PROSPECTS

1. SYNTHESIS

1.1 GENOMIC SIGNATURES

The research about mutational signatures is very active and in rapid development both in terms of new methods to analyze cancer genomic sequences and extract mutational signatures and in terms of the application of such methods with the aim to elucidate the etiology of cancer. An interesting example of current projects based on the application of mutational signatures is the Mutograph project (https://www.mutographs.org) funded by a major grant from CRUK and coordinated by the Sanger Institute in Cambridge, UK and the International Agency for Research on Cancer in Lyon, France.

This ambitious project aims to greatly extend our knowledge of the causes of several cancer types including bladder, colorectal, esophageal and kidney cancer by collecting and sequencing thousands of tumour samples, extracting the corresponding mutational signatures and link them to epidemiological data that will be collected from the participating patients. In parallel to such large applied projects, methodological research has grown extensively with an increasing number of methods to identify mutational signatures published in recent years and preprints about new methods regularly published on bioRxiv.org; we have focused on this extensive methodological work to produce a systematic review of the methods available at the time of the submission of our article, assess them and formally compare their performance.

The results of our study can be helpful to guide researchers through the planning of mutational signature analysis and provide a more solid methodological base for current projects such as Mutograph and future ones that are currently being planned. In particular, we showed that the performance of de novo methods depends on the complexity of the analyzed sequences, the number of mutations and to a lesser degree the number of samples analyzed. It was somehow expected that the performance of the methods for a cancer in which multiple, concomitant, signatures are present is poorer than for a cancer with a single or predominant signature, particularly when the concomitant signatures are similar and have a low contribution.

Additionally, we introduced a new simulation model of mutational signature data based on the zero-inflated Poisson distribution that allows for sparse contribution of signatures and thus makes it possible to build mutation count data that are more realistic than the pure Poisson model previously considered[13,151]. Finally, we improve the implementation of one of the most popular methods for signature refitting. Our method, called MutationalCone, proved to be the fastest refitting tools available to date.

1.2 EPIGENOMIC SIGNATURES

On September 3rd, 2019, Santé Publique France (SPF), the national public health agency in France published the results of a biomonitoring studies related to the presence of around 70 biomarkers including bisphenols, phthalates, BFRs, PFASs and others endocrine disruptors in the body of French citizens (Esteban, 2014-2016).

As revealed by these studies, these pollutants are omnipresent in the body of children and adults, with levels higher in the former than in the latter, findings that could be explained by dust ingestion or a high level of exposition in comparison to the body max.

With regards to BDE-47, one of the most predominant BFRs congeners observed in wildlife, mean concentrations (0.24 ng/g of lipids, N = 742) in selected the population was below the one observed in our study (0.843ng/g of lipids, N = 168). For PFOA and PFOS, observed mean were respectively, 2.08µg/L and 4.03µg/L for 744 adults aged from 18 to 74. As for BFRs, these values of SPF were below the one observed in our study. However, we should point out that our samples represent only a subset of women and their blood samples were collected in the 90s, almost 10 years before the Stockholm Convention and the associated regulations related to these compounds.

More generally, their studies reinforce the need of characterization of EDCs health 'impact. The aim of our study was to identify potential novel methylation markers of exposure to BFRs and PFASs; however we did not find evidence of moderate or strong associations between the two classes of EDCs that we investigated and methylation of DNA from blood neither at the global, genome-wide levels, at regional level (e.g. promoter regions or CpG islands) or at the level of single CpGs.

The suggestive evidence of alterations in the methylation of genes in specific biological pathways, some with plausible links with the known biological activity of PFAS and BFRs warrant further investigations in independent studies.

2. RESEARCH PERSPECTIVES

2.1 GENOMIC SIGNATURES

As argued in the section about simulated data, the simulation model that we proposed underrepresents the few samples with extremely large total mutation counts. Because catalogues of this type might hamper the detection of signals from less mutated samples in the same dataset[173], it is likely that our results slightly overestimate the methods' performance in the presence of hypermutated samples. However, our main objective was the comparison of the different methods, and this is not affected by this systematic bias. In our model, a larger number of hypermutated catalogues could be obtained by lowering the value of π, the parameter that controls the relative frequency of structural zeroes in the zero-inflated Poisson model.

As discussed, it would have been possible to consider even more realistic models, however these would have led to results that depend on too many parameters thus making the interpretation harder. For instance, the zero inflated negative binomial model is a more flexible model and looks a promising method to build realistic synthetic samples, including hypermutated ones. We leave this interesting perspective to future work. Alternative models that were recently proposed are based on the negative binomial distribution[159] and on the Dirichlet distributions for the exposures and signatures and the multinomial distribution for the catalogues[166].

We suggest that developers should assess their new methods on simulations based on realistic models such as ours or the latter. The advantage of simulations over real data is that the underlying model generating the synthetic data is known and can be compared to the estimation provided by the method being evaluated. For this reason, we decided not to simulate catalogues from real data using the bootstrap: this would have produced almost real samples but without the possibility to evaluate the performance of methods according to different parametric scenarios. We strongly believe that the mutational signature research could benefit from the development of public realistic datasets that can be used to benchmark old and new detection tools, our model is a first step in this direction.

2.2 EPIGENOMIC SIGNATURES

Given the limited sample size of in our study, it would be interesting to include additional data to study the relation between DNA methylation and BFRs or PFAS. In our work we avoided selection bias by considering only controls data from a case-control breast cancer study nested in the E3N cohort. One possibility that would allow to gain power would be to fully exploit the available data by including case data as well. In this case, selection bias could be avoided by carefully weighing cases and controls in order to have a more representative sample of the population. A formal a weighting scheme has been

recently proposed in the biostatistics literature in the context of the linear model[199], careful methodological consideration will be necessary before applying it to the mixed-effects models we used.

As BFRs and PFAS they are not the only compounds that may disrupt the endocrine system, comparative studies related to others well characterized compounds such as phthalates or bisphenol are needed, mainly to identify methylation markers involved in EDCs exposure. Additionally, other exposures, such as indoor air or dust may be considered and explored, mainly for children who are more vulnerable.

Transgenerational cohorts such as the extension to the E4N cohort that is being established with the recruitment of children and grandchildren of the E3N women, will offer an interesting opportunity to study various relationships within families that share common genetics and environments. Some studies suggest the presence of gender differences with regards to PFASs exposition[200,201].Then the effect of the dietary pattern and source of exposition to EDCs could be analyzed in the partners or offspring of E3N women to determine the concordance.

3. IMPLICATION IN PUBLIC HEALTH

As chronic diseases were becoming the leading causes of death by the middle 20th century, large-scale epidemiological studies were created to elucidate the aetiology of these diseases. Over more than half a century of research on the epidemiology of chronic diseases, has allowed to acquire extensive knowledge about why such diseases, to develop and identify their leading causes. Despite such major advances, the aetiology of several cancer types remains elusive and the recent debate about cancer and "bad luck" and the misunderstandings about the (large) extent of preventability of cancer risk to undermine the effort and achievements of several decades of epidemiological research.

Additionally, environmental exposures such as new chemicals introduced by the industry continue to emerge, contaminate and accumulate in the environment: that may pose risks related to chronic disease. These challenges require novel approaches that may take advantage of recent major advances in the analyses of biological samples with technologies such as DNA sequencing and "omics" (e.g. microarrays). It is becoming increasingly evident that environmental exposures and factors related to lifestyle may leave molecular fingerprints in various tissues that may be detected when adequate biological samples and relevant technology are available and that may provide meaningful information about the role of such factors on chronic diseases. Our findings are consistent with this general assumption and provide general support for the usefulness of studying molecular signatures to shed light on poorly understood or misunderstood aspects of cancer etiology.

We have shown for example that in realistic scenarios and under certain conditions most available methods to extract mutational signatures can accurately identify mutational signatures. With such, we have produced information that is going to be useful to guide the choice of analytical tools in important projects such as the landmark international consortium Mutographs that aims to "uncover some of the unknown causes of cancer through tell-tale signatures in DNA. Through the use of mutational signatures, we have also contributed to clarify some controversial aspects of cancer aetiology (i.e. the relative role of modifiable factors and chance) to which even the lay public has been exposed in recent years.

In addition to highlighting the potential of such novel molecular approaches to the study of chronic disease epidemiology, our work has contributed also to identify some limitations of such approach. Our analytical work on the methods of detection of mutational signatures, for example, has shown that there are scenarios for which it may be difficult to detect some of the signatures. One of these scenarios is when a tumour includes several mutational signatures each with a small contribution.

The public health implications of the results of our study on the two classes of endocrine disruptors may be difficult, partly because of the limited sample size and the possibility that such contaminants do not act through methylome changes or because blood may not be the target tissue of such action.

However, the suggestive evidence of methylome alterations in some key biological pathways, if confirmed in independent studies, may contribute to uncover potential effects on public health previously unknown or only suspected.

APPENDICES

Appendix 1. **Résumé en français**

1. Introduction 171

1.1 Les signatures mutationnelles 171

1.2 La méthylation de l'ADN 172

1.3 Les polluants organiques persistants 172

1.4 Objectifs 172

2. Matériels et méthodes 174

2.1 Identification des signatures mutationnelles 174

2.1.1 Aperçu des méthodes existantes 174

2.1.2 Simulation d'un catalogue mutationnel 174

2.1.3 La base de données TCGA 175

2.1.4 Évaluation de la performance des méthodes 175

2.2 Association entre pertubateurs endocriniens et méthylation de l'ADN 175

2.2.1 La cohorte E3N 175

2.2.2 Collection des données 175

2.2.3 Mesure du niveau circulants des BFRs et des PFASs 176

2.2.4 Méthylation de l'ADN 176

2.2.5 Gene Set Enrichment Analysis 176

2.2.6 Analyses statistiques 176

3. Résultats 177

3.1 Expositions environmentales associées aux signatures moléculaires 177

3.1.1 Expositions environmentales associées aux signatures mutationnelles et épigénétiques 177

3.1.2 Tabagisme, cancer du poumon et la role de la chance dans le dévelopement du cancer . 177

3.2 Performance des algorithmes d'identification des signatures mutationnelles 178

3.3 Association entre pertubateurs endocriniens et méthylation de l'ADN 179

3.3.1 Association entre BFRs et méthylation de l'ADN 179

3.3.2 Association entre PFASs et méthylation de l'ADN 179

4. Discussion et conclusion 180

4.1 Expositions environmentales associées aux signatures moléculaires 180

4.2 Performance des algorithmes d'identification des signatures mutationnelles 180

4.3 Association entre pertubateurs endocriniens et méthylation de l'ADN 181

1. INTRODUCTION

La compréhension des mécanismes à l'origine du développement d'un cancer ou toute autre maladie multifactorielle est essentielle pour améliorer les stratégies de prévention. À ce jour, plusieurs études ont estimé que 40% des cas de cancers observés dans les pays développés peuvent être évités en considérant les facteurs de risques connus. De même, la communauté scientifique reconnait que les expositions environnementales et le mode de vie peuvent laisser des empreintes sur l'ADN (mutations et modifications épigénétiques). Le profil mutationnel et épigénétique d'un génome résulte respectivement de la superposition de toutes les traces, ou signatures, laissées par des processus mutationnels et l'altération des niveaux de méthylation due à des facteurs environnementaux et liés au mode de vie (et à des facteurs aléatoires). La nature des données épigénétiques et génomiques étant différente (par exemple, la méthylation de l'ADN est une variable continue), des modèles mathématiques spécifiques sont nécessaires pour étudier ces deux types de signatures. Ainsi, au cours de ma thèse, j'ai étudié les approches statistiques permettant d'identifier les signatures mutationnelles ; et l'impact des perturbateurs endocriniens dans les altérations épigénétiques, un travail nécessaire pour répondre comprendre et caractériser l'effet des perturbateurs endocriniens sur la méthylation et plus globalement, sur la santé.

1.1 LES SIGNATURES MUTATIONNELLES

Les cancers résultent de diverses modifications de l'ADN comme le single nucleotide variants (ou SNV, à ne pas confondre avec SNP), insertions/délétions (ou indels), etc. ; qui se produisent généralement pendant de longues années et qui sont par la suite visibles dans l'ADN des cellules cancéreuses. Une signature génomique (ou mutationnelle) généralement notée P est définie comme étant une distribution de probabilité sur un domaine de types de mutation présélectionnés. Le domaine le plus utilisé est constitué de 96 substitutions (K=96), en considérant uniquement un nucléotide de part et d'autre de la base mutée, on parle alors de trinucléotide.

De même, au cours de son développement, un génome cancéreux *g* est exposé à différents processus mutationnels à diverses intensités. Cela se traduit par un vecteur d'exposition E dont les entrées correspondent au nombre de mutations causées en *g* par chaque signature mutationnelle n. Le catalogue mutationnel M de *g* peut alors être vu comme une superposition linéaire des n signatures avec des poids donnés par les entrées du vecteur d'exposition E, celles-ci étant généralement représentatives d'exposition environnementale.

À ce jour, plus de trente signatures mutationnelles caractérisées par un profile unique des 96 types de mutations ont été identifiées et référencées dans la base de données COSMIC (http://cancer.sanger.ac.uk/cosmic/signatures).

1.2 LA METHYLATION DE L'ADN

L'épigénome représente l'ensemble des mécanismes moléculaires impliqués dans la régulation de l'expression des gènes qui peut être influencée par l'environnement ou le mode de vie sans altération de la séquence d'ADN. Par exemple, il existe une association entre la méthylation (qui varie selon le statut tabagique) de certaines cytosines et le risque de cancer du poumon[53]. D'un point de vue moléculaire, la méthylation de l'ADN (l'une des marques épigénétique) consiste à l'ajout d'un groupement méthyl (-CH_3) sur un substrat, généralement une cytosine C. Au cours de ces dernières décennies, plusieurs découvertes ont été faites sur la méthylation de l'ADN et son importance pour un certain nombre de processus cellulaires ou de développement tels que le développement embryonnaire, l'inactivation des chromosomes X ou encore la carcinogenèse.

À titre d'exemple, les études portant sur les mécanismes moléculaires sous-jacents au rôle de la méthylation de l'ADN dans l'expression des gènes ont démontré comment les modifications épigénétiques modulent le site de liaison des facteurs de transcription à l'ADN dans les mécanismes d'activation ou d'inhibition de la transcription des gènes, et donc de la synthèse des protéines associées.

1.3 LES POLLUANTS ORGANIQUES PERSISTANTS

Les perturbateurs endocriniens sont des substances exogènes qui altèrent la ou les fonctions du système endocrinien, entraînant des effets néfastes sur la santé d'un organisme, voire de sa descendance. Cette large classe de produits chimiques comprend une variété de substances présentes dans des composants tels que les solvants industriels, les emballages alimentaires et les produits ménagers commerciaux. Leur effet sur les systèmes biologiques et leur présence répandue dans l'environnement, y compris dans les aliments, ont suscité des préoccupations croissantes quant à l'impact de leur exposition sur la santé des populations dans les pays industrialisés.

Dans ce projet de recherche, nous nous concentrerons sur les retardateurs de flamme bromés (BFRs) et les substances perfluoroalkylées et polyfluoroalkylées (PFASs), deux familles de composés connues pour perturber le système endocrinien et classées comme polluants organiques persistants, de par leur capacité à persister dans l'environnement pendant une longue période et du risque accru qu'elles représentent pour la santé humaine.

1.4 OBJECTIFS

Après l'introduction en 2013, du Framework définissant et contextualisant une signature mutationnelle, plusieurs modèles mathématiques et outils informatiques ont été proposés pour les détecter et estimer leur contribution à un catalogue donné, de même que leur association potentielle à une exposition endogène ou exogène. Ce projet avait pour objectif (1) d'examiner les contributions des signatures

mutationnelles et épigénétiques dans la conduite d'études épidémiologiques ; ce qui nous a également permis de (2) démontrer que les mutations induites par le tabagisme peuvent prédire le risque de certains cancers associés. Par la suite, (3) nous avons effectué une comparaison empirique sur la performance des outils développés pour l'analyse des signatures mutationnelles afin d'évaluer les méthodes existantes. Cela a demandé le développement d'un modèle probabiliste pour la simulation de catalogues mutationnelles réalistes sur lesquels évaluer les méthodes existantes.

Dans un second temps, et en considération de la littérature qui suggère que la méthylation joue un rôle médiateur résultant des effets des perturbateurs endocriniens sur la santé, nous nous sommes intéressés à leur potentielle association avec la méthylation de l'ADN. Nous avons donc conduit deux études afin de déterminer si la méthylation pouvait être utilisée comme biomarqueur de l'exposition aux BFRs (4) puis aux PFASs (5) en utilisant les estimations alimentaires et les mesures sanguines obtenues à partir d'une sous-population de l'étude E3N.

2. MATERIELS ET METHODES

2.1 IDENTIFICATION DES SIGNATURES MUTATIONNELLES

2.1.1 APERÇU DES METHODES EXISTANTES

La plupart des outils développés pour l'identification des signatures mutationnelles sont basées sur l'algorithme NMF[16,152] (non-negative matrix factorization) ou une version bayésienne[151,154] de celui-ci. D'autres outils sont basés sur des modèles probabilistes tels que l'algorithme EM[13]. L'objectif de toutes ces méthodes est de décomposer un catalogue mutationnel M en deux matrices P et E, dont les entrées sont non-nulles et non-négatives (à l'exception des méthodes utilisant l'ACP[152]); les signatures mutationnelles résultant des colonnes de la matrice P peuvent être alors comparées à celles référencées dans la base données COSMIC. On parle alors d'approches *de novo*.

En plus de vouloir identifier de nouvelles signatures mutationnelles, les scientifiques peuvent avoir pour intérêt, l'identification de signatures déjà existantes. On parle alors d'approches de *refitting*, qui regroupe un ensemble d'outils dont l'objectif est de trouver la meilleure combinaison de toutes les signatures existantes pouvant expliquer un catalogue mutationnel.
À ce jour, seul un modèle a été développé afin de combiner les deux approches[166].

Notre implémentation d'une méthode de refitting : MutationalCone

Dans le contexte des méthodes de refitting, nous proposons une implémentation alternative de la méthode proposée par Huang[162] ou Huebschmann[164] sur la base d'un cadre géométrique simple. En effet, trouver la décomposition linéaire du catalogue en entrée sur un ensemble de signatures données de façon à minimiser la distance entre le catalogue et une telle combinaison linéaire peut être vu comme le problème de projection sur le cône géométrique dont les arrêtes sont les signatures de référence. Nous proposons de résoudre ce problème en appliquant le package R nommé coneproj[172]. Les détails de l'implémentation de cet algorithme (MutationalCone), ainsi que le code R correspondant se trouvent dans l'Annexe 2.

2.1.2 SIMULATION D'UN CATALOGUE MUTATIONNEL

En parallèle, nous proposons également un modèle de simulation en partant du principe que le nombre de mutations induites suit une distribution de type Zero-inflated Poisson (ZIP). L'avantage de ce modèle est qu'il autorise un nombre important d'entrées nulles, ce qui correspond mieux à une modélisation hétérogène dans laquelle tous les échantillons ne sont pas exposés aux mêmes processus.

2.1.3 LA BASE DE DONNEES TCGA

En plus d'évaluer les méthodes avec des données simulées, nous avons utilisés des données réelles d'exomes de la base de données TCGA pour 4 types de pathologies : cancers du sein et du poumon, lymphome et mélanome.

2.1.4 ÉVALUATION DE LA PERFORMANCE DES METHODES

Toutes les méthodes d'identification de signatures ont pour objectif de minimiser la distance entre le catalogue réel et le produit résultant de sa décomposition. Dans un premier temps, et en utilisant les données réelles, on peut se baser sur l'erreur de reconstruction en calculant la norme de Frobenius de la différence entre la matrice avec le catalogue en entrée M et la reconstruction PxE. Par la suite, nous proposons de calculer des mesures telles que la sensibilité et la spécificité en comparant les signatures utilisées pour des simulations et celles obtenues avec les approches *de novo*. Nous simulons alors des données selon des différents valeurs de paramètres tels que le nombre de mutations et le nombre d'échantillons dans un catalogue ; les catalogues étant simulés de façon à ressembler aux catalogues des cancers sélectionnés dans TCGA. Enfin, pour évaluer les méthodes dites de *refitting*, nous comparons les biais obtenus par les méthodes en comparant l'estimation de la contribution d'une signature avec sa contribution réelle. Les méthodes sont également évaluées à l'égard du temps de calcul.

2.2 ASSOCIATION ENTRE PERTUBATEURS ENDOCRINIENS ET METHYLATION DE L'ADN

2.2.1 LA COHORTE E3N

E3N (Étude Épidémiologique auprès de femmes de la Mutuelle Générale de l'Education Nationale *(MGEN)),* est une cohorte prospective de 98 995 femmes assurées par la MGEN, dans le cadre d'un programme national de l'assurance maladie. Initiée en 1990, elle a pour objectif principal d'examiner les associations entre la mode de vie et les facteurs hormonaux, et génétiques avec le cancer et les autres maladies non-transmissibles.

2.2.2 COLLECTION DES DONNEES

Des auto-questionnaires (Q1-Q11) sont envoyés aux participantes tous les 2-3 ans afin de collecter les données relatives à leur état de santé et mode de vie. Il existe également une banque biologique constituée avec des échantillons sanguins collectés entre 1994 et 1999 chez environ 25 000 participantes (taux de participation ~40%) et salivaires collectés entre 2009 et 2011 chez 47 000 femmes (taux de participation ~70%). De plus, les données de la MGEN sont disponibles depuis 2004 et fournissent des informations sur les remboursements des médicaments des femmes E3N.

Les données alimentaires sont disponibles grâce à deux questionnaires portant sur les habitudes alimentaires des années antérieures envoyés en 1993 et en 2005. L'estimation de l'exposition alimentaire aux BFRs et aux PFASs ont par la suite été basées sur le questionnaire alimentaire envoyé en 1993 en utilisant la base de données TDS2 qui regroupe plus de 20 000 produits alimentaires servant de support pour l'identification de 1352 composés.

2.2.3 MESURE DU NIVEAU CIRCULANTS DES BFRS ET DES PFASS

Les niveaux circulants des BFRs et des PFASs d'environ 200 cas et 200 témoins du cancer du sein ont été mesurés par le laboratoire LABERCA (Oniris Nantes, FRANCE) en utilisant les protocoles adaptés selon la norme ISO. Ces femmes ont été appariés sur la base de l'âge, l'IMC, le statut ménopausique et le département de résidence au prélèvement sanguin.

2.2.4 METHYLATION DE L'ADN

La puce illumina HumanMethylation EPIC a été utilisée pour mesurer le niveau de plus de 850 K CpGs le long du génome. L'extraction de l'ADN, le protocole de conversion, le contrôle qualité et le prétraitement ont été réalisé par l'Italian Institute of Genomic Medicine (IIGM).

2.2.5 GENE SET ENRICHMENT ANALYSIS

GSEA[185] est une méthode qui permet de déterminer si un ensemble de gènes présente des différences concordantes avec un état biologique, ex. un phénotype binaire ou une corrélation avec un phénotype quantitatif, ex. niveau de BFRs. Dans le cadre de cette thèse, il s'agit d'évaluer la corrélation entre la méthylation des CpGs localisés dans les régions promotrices de gènes et le niveau des différents perturbateurs endocriniens étudiés.

2.2.6 ANALYSES STATISTIQUES

Dans le cadre de cette thèse, seuls les témoins de l'enquête cas/témoins nichée dans la cohorte ont été considérés afin d'éviter le biais de sélection consistant à étudier l'association entre méthylation et exposition conditionnellement au statut cas/contrôle, un effet potentiellement commun à ces deux variables. L'association entre BFRs ou PFASs et méthylation de l'ADN a été évalué à l'égard des CpGs pris individuellement, de leur niveau moyen sur des régions, et du niveau moyen global.
Les populations finales de l'étude sur l'association entre les BFRs et les PFASs et la méthylation de l'ADN variaient entre 162 et 168 femmes. Les quartiles d'exposition alimentaires ou de mesures des niveaux circulants ont été étudiés en relation avec niveau de méthylation de l'ADN en utilisant divers modèles linéaires à effet mixtes. En fonction du modèle, les facteurs d'ajustements prenaient en compte l'âge, l'IMC, la parité/durée cumulée d'allaitement, le score d'adhérence au régime alimentaire méditerranéen ou occidental et le taux total de lipides.

3. RESULTATS

3.1 EXPOSITIONS ENVIRONMENTALES ASSOCIEES AUX SIGNATURES MOLECULAIRES

3.1.1 EXPOSITIONS ENVIRONMENTALES ASSOCIEES AUX SIGNATURES MUTATIONNELLES ET EPIGENETIQUES

Dans un premier temps, nous avons effectué une revue de littérature portant sur les associations connues entre exposition environnementales et signatures mutationnelles ou épigénétiques.

Les rayons ultraviolets (UV) sont connus pour induire des transitions de type C > T tandis que le tabagisme induit majoritairement des transversions C > A. Cela a pu être prouvé expérimentalement, et le principe selon lequel les carcinogènes laissent des empreintes a pu être confirmé avec la disponibilité des données d'exomes et de génomes de multiples cancers.

À titre d'exemple, une étude[118] portant sur 2490 fumeurs et 1063 non-fumeurs, a permis d'identifier une prévalence plus importante de signatures mutationnelles chez les fumeurs en comparaisons des non-fumeurs. De même, elle a permis d'identifier la signature mutationnelle 4 comme résultant de l'exposition au tabagisme avec une fraction considérable chez les fumeurs pour les cancers du poumon du larynx et du foie ; la signature 4 étant majoritairement constituée de transversions C>A.

Plus généralement, en termes d'association entre les signatures mutationnelles et les expositions environnementales, la plupart des autres études portent sur l'acide aristolochique[127], l'aflatoxine B1[120], les rayons UV[11] et les radiations[121,122].

Par ailleurs, il existe une association entre la méthylation (qui varie selon le statut tabagique) de certaines cytosines et le risque de cancer du poumon. Il a par exemple été démontré qu'il existe des différences supérieures à 5% entre le niveau de méthylation dans les tissus tumoraux de fumeurs en comparaison des non-fumeurs[118].

3.1.2 TABAGISME, CANCER DU POUMON ET LA ROLE DE LA CHANCE DANS LE DEVELOPEMENT DU CANCER

Le cancer du poumon est le 3ème cancer au monde et plusieurs études ont permis de démontrer que le tabagisme en est la cause principale.

Depuis 2015, Tomasetti et Vogelstein ont publiés un certain nombre d'articles qui ont contribué du fait d'une certaine ambiguïté, à la diffusion de l'idée, fausse, selon laquelle 2/3 des nouveaux cas de cancers résulteraient du hasard. Leur modèle, qui se base notamment sur l'estimation du nombre de divisons de cellules souches (LSCD), leur a permis de déterminer qu'aux États-Unis, mais également dans 68 autres pays, le LSCD de 25 types de tissus corrèle bien avec le risque de développer un cancer (CR) dans ces mêmes tissus. En particulier, la variation du log (CR) expliquée par le log (LSCD) serait de $R^2 = 0.66$.

Bien que cette mesure de corrélation ne soit déterminante sur la probabilité de développer un cancer, elle a été interprétée comme une mesure de la proportion de nouveaux cancers dus à la malchance. En 2017, les auteurs ont clarifié leurs propos en effectuant une distinction claire entre la proportion de cancers évitables dus à l'exposition environnementale et la proportion de mutations déterminantes causées par des facteurs environnementaux, l'hérédité ou des facteurs stochastiques incontrôlables (notamment les erreurs lors de la réplication de l'ADN). Cependant, ce modèle présente également des failles puisque les mutations sont nécessaires ; mais pas suffisantes pour aboutir au développement du cancer.

Certains chercheurs ont proposé une alternative pour estimer le nombre de mutations due à des facteurs endogènes ou exogènes en se basant sur les signatures mutationnelles[147]. Étant donné que la signature COSMIC 1 est corrélée à l'âge de diagnostic du cancer, ses chercheurs ont utilisé le ratio entre le nombre de mutations associées à cette signature et le nombre total de mutations totale en tant proxy de la proportion de mutations intrinsèques. En utilisant cette approche, ils ont estimé que la grande majorité des mutations (70% à 90%) est due à des facteurs extrinsèques dans la plupart des types de cancer, ce qui contredit les conclusions de Tomasetti et Vogelstein. En utilisant une approche similaire, nous avons comparé le nombre le nombre de mutations dues au tabagisme dans plusieurs tissus, à l'incidence et au taux de mortalité de multiples cancers associés au tabagisme chez les fumeurs et les non-fumeurs.

Nos résultats démontrent ainsi que le nombre de mutations est plus prédictif du risque de cancer que le nombre de divisions cellulaires.

3.2 PERFORMANCE DES ALGORITHMES D'IDENTIFICATION DES SIGNATURES MUTATIONNELLES

En considérant les données réelles, on remarque que l'erreur de reconstruction des différentes méthodes dépend du type de cancer, ce qui est attendu puisque ces jeux de données varient à l'égard du nombre d'échantillons, de mutations et du nombre de signatures. Plus généralement, toutes les méthodes sont rigoureusement équivalentes dans leur capacité à reconstruire le catalogue mutationnel initial.

Nous n'observons pas de larges différences pour la spécificité à l'égard du nombre d'échantillon dans le catalogue. Il faut cependant noter que la sensibilité augmente avec le nombre d'échantillons ; notamment dans le cas où plusieurs signatures contribuent au profil mutationnel d'un catalogue. Par ailleurs, les méthodes basées sur la NMF ont une sensibilité moindre tandis que celles basées sur les modèles probabilistes donnent de meilleurs résultats.
La sensibilité augmente avec le nombre de mutations et pour la majorité des cas, une moyenne de 1000 mutations sont nécessaires pour qu'elle avoisine 1. Plus généralement, elle est élevée pour les cancers

ayant une mutation prédominante et faible en cas de signature concomitantes. Des résultats comparables sont observés pour la spécificité.

En simulant un catalogue de cancer du poumon, et en appliquant les méthodes de refitting, nous nous apercevons qu'elles donnent de bonnes estimations de la contribution de la majorité des signatures. En termes de temps de calcul, l'algorithme que nous proposons, MutationalCone s'avère être le plus rapide.

3.3 ASSOCIATION ENTRE PERTUBATEURS ENDOCRINIENS ET METHYLATION DE L'ADN

Les analyses individuelles CpG par CpG et les analyses de la méthylation de l'ADN aux niveaux global et régional n'ont pas fourni de preuve convaincante d'associations avec les concentrations plasmatiques des BFRs/PFASs ou l'exposition alimentaire aux BFRs/PFASs. Les résultats de l'analyse GSEA sont tout de même intéressants car ils suggèrent que l'exposition aux BFRs ou aux PFASs peuvent modifier les niveaux de méthylation de l'ADN de gènes impliqués dans des voies biologiques spécifiques.

3.3.1 ASSOCIATION ENTRE BFRS ET METHYLATION DE L'ADN

Les concentrations plasmatiques et l'exposition alimentaire aux BFRs semblent être associées à des altérations de la méthylation de l'ADN dans différentes voies métaboliques. Tandis que pour les niveaux circulants, les groupes de gènes identifiés sont impliqués dans l'embryogénèse, la régulation de la matrice extracellulaire et du cycle cellulaire et les mécanismes de réparation de l'ADN, les expositions alimentaires sont associées à des voies telles que la réponse immunitaire, à l'hypoxie et à l'apoptose.

Ces résultats sont globalement compatibles avec la capacité des BFRs à modifier le système endocrinien, influencer la réponse immunitaire et impacter le système reproducteur.

3.3.2 ASSOCIATION ENTRE PFASS ET METHYLATION DE L'ADN

Tout comme les BFRs, les concentrations plasmatiques et l'exposition alimentaire aux PFASs semblent être associées à des altérations de la méthylation de l'ADN dans différentes voies. Tandis que pour les niveaux circulants, les groupes de gènes identifiés sont impliqués dans la régulation de l'homostase du cholestérol et le métabolisme des acides gras, les expositions alimentaires sont principalement associées à l'apoptose.

Ces résultats sont globalement compatibles avec la capacité des PFASs à influencer la réponse immunitaire et à leurs propriétés hépatotoxiques.

4. DISCUSSION ET CONCLUSION

4.1 EXPOSITIONS ENVIRONMENTALES ASSOCIEES AUX SIGNATURES MOLECULAIRES

Le profil mutationnel et épigénétique d'un génome cancéreux résulte respectivement de la superposition de toutes les traces, ou signatures, laissées par des processus mutationnels et de l'altération des niveaux de méthylation dues à des facteurs environnementaux, de style de vie (et aléatoires). Ces deux types de signatures représentent des domaines de recherche prometteurs susceptibles de continuer à apporter de nouvelles connaissances sur la nature du cancer et les processus qui y conduisent. Ces avancées dans les nouvelles connaissances vont probablement s'accélérer lorsque des études épidémiologiques vont collecter et séquencer systématiquement l'ADN du tissu tumoral, permettant ainsi l'analyse des signatures mutationnelles et la mise en relation de ces signatures avec des données épidémiologiques.
Selon le modèle dominant de cancérogenèse, le cancer est principalement causé par l'accumulation de mutations génétiques. Cependant, il est de plus en plus admis que l'accumulation de mutations somatiques ne peut à elle seule expliquer le développement d'un cancer. Les preuves s'accumulent et il est reconnu que les mécanismes génétiques ou non génétiques tels que les altérations épigénétiques et les facteurs environnementaux peuvent influencer les divisions des cellules souches et donc le développement du cancer. À cet égard, il serait très intéressant d'essayer d'estimer l'effet de tels facteurs sur le nombre de divisions de cellules souches au cours de la vie. Cela nécessiterait la construction d'un modèle permettant d'estimer la fraction de tels événements par rapport au nombre total d'événements nécessaires au développement du cancer. D'autres événements ou conditions pouvant jouer un rôle important mais qui n'ont pas encore été pris en compte dans le modèle de développement du cancer sont les mécanismes de réparation de l'ADN et les dysfonctionnements de la surveillance immunitaire.

4.2 PERFORMANCE DES ALGORITHMES D'IDENTIFICATION DES SIGNATURES MUTATIONNELLES

La recherche sur les signatures mutationnelles est très active et se développe rapidement, à la fois en ce qui concerne les nouvelles méthodes d'analyse des séquences génomiques du cancer, mais également, l'application de ces méthodes dans le but d'élucider l'étiologie du cancer.

Les résultats des travaux menés portant sur la comparaison des méthodes d'identification des signatures mutationnelles permettent de mieux comprendre les forces et les limites de chaque méthode, ainsi que l'identification les paramètres clés qui influent leurs performances, à savoir le nombre de mutations et la « complexité » des facteurs contributifs, notamment les signatures.

De même, notre étude semble indiquer que les méthodes probabilistes de novo EMu et bayesNMF ont globalement une meilleure performance car elles permettent d'obtenir une sensibilité et une spécificité

meilleures avec un temps de calcul raisonnable. Cependant, afin d'évaluer la robustesse des nouveaux résultats, en raison de la variabilité des résultats et de la présence d'échantillons hypermutés notamment, nous recommandons d'effectuer systématiquement une analyse de sensibilité basée sur l'application d'une ou de plusieurs méthodes alternatives basées sur différents algorithmes.

Plus généralement, si l'objectif est d'évaluer la présence de signatures connues dans des catalogues de mutations (génomes ou exomes de cancer), nous recommandons de passer aux méthodes de *refitting*. Pour les cancers bien étudiés, elles constituent une alternative plus rapide et plus puissante que les méthodes de novo. Comme la base de données COSMIC a été construite et validée en analysant des dizaines de milliers de séquences de la plupart des types de cancer, il est recommandé de s'appuyer sur les études précédentes et d'utiliser des outils de *refitting* pour réaliser une analyse standard ne visant pas la découverte de signatures *de novo*.

Par ailleurs, nous avons introduit un nouveau modèle de simulation de données de signatures mutationnelles basé sur une distribution de Poisson (ZIP) qui permets d'obtenir des simulations plus réalistes que celles récemment proposées dans la littérature. De même, nous avons proposé une version améliorée des modèles de *refitting* existants, et notre méthode, appelée MutationalCone, s'est révélée être l'outil de ce type le plus rapide disponible à ce jour.

4.3 ASSOCIATION ENTRE PERTUBATEURS ENDOCRINIENS ET METHYLATION DE L'ADN

Le 3 septembre 2019, Santé Publique France, l'agence nationale de santé publique en France, a publié les résultats d'une étude de biosurveillance liée à la présence d'environ 70 biomarqueurs, notamment des bisphénols, des phtalates, des BFRs, des PFASs et d'autres perturbateurs endocriniens dans l'organisme des Français (Esteban, 2014-2016).Comme le révèle l'étude, ces polluants sont omniprésents dans le corps des enfants et des adultes, avec des niveaux plus élevés chez les enfants, ce qui pourrait s'expliquer par l'ingestion de poussière ou par un niveau d'exposition élevé par rapport au poids de leur corps. Cela ne fait donc que refléter l'importance d'étudier l'impact de telles molécules sur la santé.

Le but de notre étude était d'identifier de nouveaux marqueurs d'exposition aux BFRs et aux PFASs. Cependant, nous n'avons trouvé aucune preuve d'association modérée ou forte entre ces deux classes de perturbateurs endocriniens et la méthylation de l'ADN sanguin, que ce soit au niveau global, à l'échelle du génome, régional (par exemple, les régions promotrices aux ilôts de CpG), ou des CpGs pris individuellement.

À notre connaissance, il s'agit de la première étude d'association à l'échelle de l'épigénome des BFRs ou des PFASs et de la méthylation de l'ADN. Les études antérieures portaient sur des éléments génomiques répétitifs utilisés comme marqueurs de la méthylation globale (à savoir Alu et LINE-1), et ont démontrés

que des perturbateurs endocriniens tels que les phtalates ou les bisphénols étaient associés à une hypométhylation. Notre étude a mesuré la méthylation de l'ADN de manière plus systématique avec une couverture de près d'un million de CpG représentant plus de 90% de tous les CpG - soit une couverture six fois supérieure à la couverture des études utilisant des éléments Alu et LINE-1.

Les principales limites de notre étude incluent la nature transversale des mesures dans le sang (c'est-à-dire que les concentrations plasmatiques des BFRs ou des PFASs et la méthylation de l'ADN ont été mesurées à partir des mêmes échantillons de sang) et la taille relativement limitée des populations étudiées. En outre, nous ne pouvons pas exclure que la possibilité que les BFRs ou les PFASs influencent la méthylation de l'ADN dans d'autres tissus non disponibles pour cette étude.

En conclusion, notre étude n'a trouvé aucune preuve d'association entre l'exposition aux BFRs ou aux PFAS et des altérations modérées ou fortes de la méthylation des CpG pris globalement ou individuellement dans l'ADN circulant. Les associations observées entre l'exposition aux BFRs ou aux PFASs et les altérations de la méthylation de l'ADN dans des voies biologiques spécifiques méritent d'être répliquées dans des études indépendantes puisqu'elles pourraient refléter une action plus complexe de cette classe de substances.

Appendix 2. **MutationalCone implementation**

We report here the R code implementing our original method for signature refitting.

Let S be the linear subspace of $\mathbb{R}^K$ spanned by the reference signatures. Our function MutationalCone() projects the input mutational catalogue onto the cone in S spanned by the reference signatures with the very fast coneproj R package (https://cran.r-project.org/web/packages/coneproj). Because projections are simply calculated as scalar products, this function requires the user to specify an orthonormal basis of S toghether with the components of the reference signatures with respect to it. These two input matrices can be calculated with the function SignatureSubspace() once and for all, before iterating MutationalCone() on all catalogues. SignatureSubspace() finds an orthonormal basis of S with the Gram-Schmidt algorithm.

```
SignatureSubspace <- function(signatures){
  # signatures: (K,N)-matrix with reference signatures in columns (e.g.     #
COSMIC signatures)
  # with K = number of mutation types (e.g. 96), N = number of reference    #
signatures

  # Orthonormalization of the subspace generated by reference signatures
  S <- signatures
  S.qr <- qr(S)
  Q <- qr.Q(S.qr) # orthonormal basis of the subspace
  R <- qr.R(S.qr) # components of the reference signatures in the orthonormal
basis
  return(list(Q=Q, R=R))
}

MutationalCone <- function(catalogue, Q, R){
  # catalogue: vector of length K with the mutational catalogue,
  # Q: matrix with the orthonormal basis of the subspace generated
  # by the reference signatures in columns
  # R: matrix with the components of the reference signatures wrt
  # the orthonormal basis in columns. Q and R are found with                #
SignatureSubspace()

  require(coneproj)

  # Projection of the catalogue onto the subspace generated by
  # reference signatures
  proj.subspace <- t(Q) %*% catalogue

  # Projection onto the cone spanned by the signatures
  weights <- as.vector(coneB(y=as.vector(proj.subspace),delta=R)$coefs)
  return(weights)
}
```

Appendix 3. **Overview of gene sets in MSigDB**
http://software.broadinstitute.org/gsea/msigdb/collections.jsp

H: hallmark gene sets (browse 50 gene sets)	Hallmark gene sets summarize and represent specific well-defined biological states or processes and display coherent expression. These gene sets were generated by a computational methodology based on identifying overlaps between gene sets in other MSigDB collections and retaining genes that display coordinate expression. details
C1: positional gene sets (browse 299 gene sets)	Gene sets corresponding to each human chromosome and each cytogenetic band that has at least one gene. details
C2: curated gene sets (browse 5501 gene sets)	Gene sets curated from various sources such as online pathway databases, the biomedical literature, and knowledge of domain experts. The gene set page for each gene set lists its source. The C2 collection is divided into two sub-collections: CGP and CP.
CGP: chemical and genetic perturbations (browse 3302 gene sets)	Gene sets represent expression signatures of genetic and chemical perturbations. A number of these gene sets come in pairs: xxx_UP (and xxx_DN) gene set representing genes induced (and repressed) by the perturbation.
CP: Canonical pathways (browse 2199 gene sets)	Gene sets from pathway databases. Usually, these gene sets are canonical representations of a biological process compiled by domain experts.
CP:BIOCARTA: BioCarta gene sets (browse 289 gene sets)	Gene sets derived from the BioCarta pathway database.
CP:KEGG: KEGG gene sets (browse 186 gene sets)	Gene sets derived from the KEGG pathway database.
CP:PID: PID gene sets (browse 196 gene sets)	Gene sets derived from the PID pathway database.
CP:REACTOME: Reactome gene sets (browse 1499 gene sets)	Gene sets derived from the Reactome pathway database.
C3: motif gene sets (browse 831 gene sets)	Gene sets representing potential targets of regulation by transcription factors or microRNAs. The sets consist of genes grouped by short sequence motifs they share in their non-protein coding regions. The motifs represent known or likely cis-regulatory elements in promoters and 3'-UTRs. The C3 collection is divided into two sub-collections: MIR and TFT details
MIR: microRNA targets (browse 221 gene sets)	Gene sets that contain genes sharing putative target sites (seed matches) of human mature miRNA in their 3'-UTRs.
TFT: transcription factor targets (browse 610 gene sets)	Gene sets that share upstream CIS-regulatory motifs which can function as potential transcription factor binding sites. Based on work by Xie et al. 2005

Continued on the following page

C4: computational gene sets (browse 858 gene sets)	Computational gene sets defined by mining large collections of cancer-oriented microarray data. The C4 collection is divided into two sub-collections: CGN and CM. details
CGN: cancer gene neighborhoods (browse 427 gene sets)	Gene sets defined by expression neighborhoods centered on 380 cancer-associated genes. This collection is described in Subramanian, Tamayo et al. 2005
CM: cancer modules (browse 431 gene sets)	Gene sets defined by Segal et al. 2004. Briefly, the authors compiled gene sets ('modules') from a variety of resources such as KEGG, GO, and others. By mining a large compendium of cancer-related microarray data, they identified 456 such modules as significantly changed in a variety of cancer conditions.
C5: GO gene sets (browse 9996 gene sets)	Gene sets that contain genes annotated by the same GO term. The C5 collection is divided into three sub-collections based on GO ontologies: BP, CC, and MF. details
BP: GO biological process (browse 7350 gene sets)	Gene sets derived from the GO Biological Process Ontology.
CC: GO cellular component (browse 1001 gene sets)	Gene sets derived from the GO Cellular Component Ontology.
MF: GO molecular function (browse 1645 gene sets)	Gene sets derived from the GO Molecular Function Ontology.
C6: oncogenic signatures (browse 189 gene sets)	Gene sets that represent signatures of cellular pathways which are often dis-regulated in cancer. The majority of signatures were generated directly from microarray data from NCBI GEO or from internal unpublished profiling experiments involving perturbation of known cancer genes. details
C7: immunologic signatures (browse 4872 gene sets)	Gene sets that represent cell states and perturbations within the immune system. The signatures were generated by manual curation of published studies in human and mouse immunology. details

Appendix 4. **Description of hallmarks associated with BFRs or PFASs exposure**

Gene set	Description
ANDROGEN_RESPONSE	Genes defining response to androgens
APOPTOSIS	Genes mediating programmed cell death (apoptosis) by activation of caspases
CHOLESTEROL_HOMEOSTASIS	Genes involved in cholesterol homeostasis
DNA_REPAIR	Genes involved in DNA repair
FATTY_ACID_METABOLISM	Genes encoding proteins involved in metabolism of fatty acids
HYPOXIA	Genes up-regulated in response to low oxygen levels (hypoxia)
IL2_STAT5_SIGNALING	Genes up-regulated by STAT5 in response to IL2 stimulation
IL6_JAK_STAT3_SIGNALING	Genes up-regulated by IL6 via STAT3 e.g. during acute phase response
INFLAMMATORY_RESPONSE	Genes defining inflammatory response
MYC_TARGETS_V1	A subgroup of genes regulated by MYC - version 1
MYC_TARGETS_V2	A subgroup of genes regulated by MYC - version 2
TNFA_SIGNALING_VIA_NFKB	Genes regulated by NF-kB in response to TNF

Appendix 5. **Top 20 CpGs associated with dietary exposure to HBCDs congeners**

	Probes	Coefficients*	SE	P
HBCDalpha	cg27595499	-0,2760762	0,05816591	6,50904E-06
	cg07390488	0,24457846	0,05346267	1,29971E-05
	cg05133593	0,26558025	0,06013522	2,42746E-05
	cg14401140	0,4976144	0,1128951	2,51051E-05
	cg00770317	0,56135797	0,12873052	3,01393E-05
	cg27661315	0,41997914	0,09668882	3,22008E-05
	cg07142556	0,29012627	0,06736854	3,71439E-05
	cg20189913	0,33286699	0,07742552	3,82096E-05
	cg02370023	0,28719499	0,06787434	4,95276E-05
	cg09345606	-0,3623646	0,08594503	5,24379E-05
	cg20320200	0,15395841	0,03653598	5,29078E-05
	cg23956068	0,36902208	0,08768005	5,39506E-05
	cg25769013	0,31804512	0,07572767	5,57955E-05
	cg13279940	0,5621625	0,13453731	6,04856E-05
	cg07787543	-0,2842825	0,0680523	6,07315E-05
	cg07884019	-0,2701917	0,06490753	6,41878E-05
	cg05169756	0,33368971	0,08026058	6,54403E-05
	cg07829740	0,34311475	0,08282472	6,92111E-05
	cg18675735	0,42777874	0,1035916	7,27175E-05
	cg04156077	2,10044806	0,5127395	8,22392E-05
HBCDbeta	cg18404184	0,3834896	0,06996678	2,88166E-07
	cg02786218	0,23479543	0,04483681	8,34626E-07
	cg00825491	0,20337156	0,03908586	9,63578E-07
	cg06019792	0,15252194	0,03014283	1,77213E-06
	cg06409164	0,1984154	0,04060276	3,65679E-06
	cg20189913	0,35113083	0,07246902	4,34053E-06
	cg01454153	-0,3997916	0,08343325	5,40751E-06
	cg19788036	0,33911074	0,07084019	5,51418E-06
	cg15267844	1,17891241	0,24750664	6,0782E-06
	cg22500518	0,15701139	0,03365453	9,03712E-06
	cg17232357	0,1589013	0,03433133	1,04843E-05
	cg06210526	-0,5255223	0,11451435	1,2272E-05
	cg03772491	0,14651673	0,03192809	1,22808E-05
	cg04695063	0,19459668	0,04257474	1,32083E-05
	cg19947484	0,18671063	0,04094203	1,37634E-05
	cg11593179	0,2341994	0,05205816	1,7566E-05
	cg11048101	0,82583464	0,18484974	1,98628E-05
	cg15032048	0,26068104	0,05836155	1,99366E-05
	cg10941185	0,20092138	0,04545865	2,39464E-05
	cg11085508	0,31127961	0,07055069	2,46798E-05

Continued on the following page

HBCDgamma	cg06409164	0,2061523	0,03818988	4,14578E-07
	cg17562250	0,29351139	0,05759655	1,52172E-06
	cg18404184	0,33956167	0,06722252	1,83814E-06
	cg08685384	0,15303902	0,03092416	2,82497E-06
	cg11948159	0,21965814	0,04656777	7,33548E-06
	cg18493250	0,24804296	0,0530769	8,75405E-06
	cg00825491	0,17494655	0,03755575	9,29836E-06
	cg15267844	1,09585877	0,23533895	9,36617E-06
	cg23806034	0,135452	0,02917427	9,89527E-06
	cg11048101	0,80641673	0,17484133	1,11873E-05
	cg22711299	0,30495955	0,0672754	1,53434E-05
	cg06384026	0,15275811	0,03396177	1,76257E-05
	cg07913620	0,25900233	0,05779446	1,88104E-05
	cg11702456	0,27860728	0,06230638	1,95553E-05
	cg19788036	0,302456	0,06765352	1,96252E-05
	cg13421489	0,04811064	0,01085596	2,28622E-05
	cg19526199	-0,4068794	0,09198272	2,36155E-05
	cg08445278	0,20454323	0,04699345	3,11029E-05
	cg25304608	0,16170754	0,03731143	3,34195E-05
	cg21662240	0,47488652	0,10983435	3,47777E-05
HBCDs	cg25769013	0,37908579	0,07699725	3,14123E-06
	cg20189913	0,37333789	0,07972669	8,42636E-06
	cg07390488	0,25767529	0,05519859	8,93697E-06
	cg12000297	0,80574196	0,17273487	9,06523E-06
	cg02370023	0,3189253	0,07005543	1,42024E-05
	cg23956068	0,41035208	0,09035225	1,48236E-05
	cg07142556	0,31570191	0,06976916	1,5841E-05
	cg05133593	0,27998831	0,06216846	1,72294E-05
	cg27595499	-0,2739161	0,06088536	1,75605E-05
	cg22989447	0,42564279	0,09468147	1,77947E-05
	cg14652403	0,20668853	0,04664839	2,29476E-05
	cg04156077	2,30773125	0,52531633	2,6587E-05
	cg07829740	0,3754783	0,08562843	2,74333E-05
	cg15267844	1,20162834	0,27564012	3,02927E-05
	cg00770317	0,58130742	0,1335935	3,12567E-05
	cg19389613	-0,2456931	0,05703641	3,69874E-05
	cg18675735	0,46033789	0,10706749	3,81624E-05
	cg04277282	0,2398269	0,05591038	3,96585E-05
	cg05169756	0,35425044	0,08330199	4,56647E-05
	cg10836258	0,27148176	0,06386742	4,5996E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern as fixed effects

Appendix 6. **Top 20 CpGs associated with dietary exposure to PBDEs congeners**

	Probes	Coefficients*	SE	P
BDE-28	cg02874371	0,63060434	0,1270389	2,65332E-06
	cg22855255	0,24835987	0,05197901	5,7187E-06
	cg02466588	0,26346077	0,0573153	1,19072E-05
	cg20136236	0,28976953	0,06328276	1,27806E-05
	cg03801924	-0,2933135	0,06479058	1,57066E-05
	cg13062913	0,20185849	0,0447471	1,67334E-05
	cg15424250	0,18542467	0,04119381	1,73954E-05
	cg24679242	0,32749441	0,07318109	1,92838E-05
	cg02447557	0,33412354	0,07533846	2,25744E-05
	cg15134456	-0,345273	0,0780331	2,3499E-05
	cg13223537	-0,2744715	0,06215535	2,4321E-05
	cg10009007	-0,3154531	0,07185485	2,68873E-05
	cg25780498	0,34592269	0,07887102	2,73314E-05
	cg17862113	0,17750777	0,04061364	2,90023E-05
	cg03507241	0,24400807	0,05607367	3,12268E-05
	cg06924602	-0,3437152	0,07899194	3,12622E-05
	cg08351563	-0,3935994	0,09087029	3,37502E-05
	cg00678890	0,25551947	0,05913703	3,51626E-05
	cg25132878	0,27326286	0,06349789	3,7583E-05
	cg19944002	-0,2673953	0,0621455	3,76937E-05
BDE-47	cg20136236	0,29821485	0,06217601	5,30773E-06
	cg19720347	0,11989333	0,02511697	5,82914E-06
	cg18538510	0,22225298	0,04792463	1,01086E-05
	cg25683662	0,23241462	0,0510195	1,40376E-05
	cg23706176	-0,2836156	0,06357924	2,03988E-05
	cg03940874	-0,3720182	0,0834923	2,08115E-05
	cg18349130	-0,2593741	0,05825947	2,11132E-05
	cg08351563	-0,3974474	0,08947593	2,19658E-05
	cg13062913	0,19536953	0,04436356	2,54925E-05
	cg25197238	0,18088281	0,04129988	2,79963E-05
	cg01102638	0,39152512	0,09009864	3,19653E-05
	cg17741837	0,57537073	0,13331069	3,5817E-05
	cg02874371	0,54652215	0,12777539	4,15681E-05
	cg02380813	0,24057125	0,0564808	4,45052E-05
	cg09285095	0,28814335	0,06769421	4,49837E-05
	cg18787229	0,21345459	0,05064749	5,27816E-05
	cg03801924	-0,2707463	0,06431054	5,3696E-05
	cg01518607	0,30928869	0,07355488	5,47462E-05
	cg17256404	0,21813041	0,05195872	5,61567E-05
	cg01383890	0,23749504	0,05659204	5,64831E-05
	cg03243551	0,29297613	0,0606766	4,65078E-06

Continued on the following page

BDE-99	cg12809314	0,35816084	0,07630751	8,06203E-06
	cg10009007	-0,3287514	0,07020614	8,42849E-06
	cg20189913	0,32939542	0,07051444	8,82344E-06
	cg22857777	0,14351029	0,03078702	9,18395E-06
	cg22788368	0,25824114	0,055488	9,46144E-06
	cg26671685	0,77246414	0,16744097	1,11393E-05
	cg05637795	0,14929357	0,03282772	1,4469E-05
	cg19486875	0,16877637	0,03720143	1,51127E-05
	cg05134987	0,48245701	0,10647776	1,54621E-05
	cg01992382	0,40500369	0,08964232	1,6283E-05
	cg16834823	0,38341855	0,0853086	1,78675E-05
	cg24015654	-0,2168261	0,04829003	1,81808E-05
	cg11712934	0,27007556	0,06020563	1,84839E-05
	cg15670585	0,12113519	0,02706113	1,91907E-05
	cg05575043	0,33734257	0,07573865	2,09495E-05
	cg25769013	0,30573859	0,06885379	2,20991E-05
	cg08439122	0,20502268	0,0462413	2,26815E-05
	cg25161161	0,33497904	0,07568143	2,33642E-05
	cg21169617	0,23813056	0,05423803	2,68529E-05
BDE-100	cg20136236	0,30023388	0,06319181	6,38312E-06
	cg22855255	0,24659666	0,05211227	6,8999E-06
	cg17080882	0,24684448	0,0540564	1,34355E-05
	cg19720347	0,11565438	0,0258721	1,96582E-05
	cg08500500	0,32711072	0,07357879	2,16436E-05
	cg14102355	0,63115141	0,1436603	2,6553E-05
	cg05764121	-0,3325359	0,0757127	2,6686E-05
	cg02874371	0,56817291	0,12936611	2,66963E-05
	cg15134456	-0,341506	0,07825177	2,97364E-05
	cg22788368	0,2451431	0,05644348	3,22576E-05
	cg15424250	0,17980455	0,04145155	3,29445E-05
	cg23706176	-0,2806779	0,06484336	3,41283E-05
	cg25780498	0,34100604	0,079297	3,80375E-05
	cg02447557	0,3257715	0,07579895	3,84086E-05
	cg10009007	-0,3090468	0,07220673	4,11221E-05
	cg21226850	-0,2888084	0,06760066	4,23609E-05
	cg25683662	0,22550677	0,05279827	4,25507E-05
	cg24714511	0,19803863	0,04677432	4,90325E-05
	cg01414857	0,27549107	0,06508573	4,92523E-05
	cg24679242	0,31203071	0,07385352	5,07237E-05

Continued on the following page

BDE-153	cg27268574	0,2090347	0,04232729	2,94911E-06
	cg15851014	0,28348095	0,05892163	4,99409E-06
	cg02884197	0,68616238	0,14309896	5,33566E-06
	cg05155449	0,23479678	0,04934279	6,19518E-06
	cg25561579	-0,4650648	0,09804318	6,58606E-06
	cg09125532	0,2768987	0,05862835	7,15919E-06
	cg03819515	0,20288978	0,04373719	1,00564E-05
	cg15842366	0,22981696	0,04966191	1,05195E-05
	cg00745323	0,5802423	0,12627952	1,19926E-05
	cg13066682	0,10391439	0,02269445	1,27867E-05
	cg03920024	0,2037169	0,0449072	1,51383E-05
	cg06384026	0,17004806	0,03752264	1,54119E-05
	cg19283806	0,19561402	0,04318171	1,55259E-05
	cg22788368	0,27158873	0,06028324	1,71271E-05
	cg27273140	0,31594558	0,07023583	1,75972E-05
	cg00770317	0,5710177	0,12702525	1,78098E-05
	cg06019792	0,14494261	0,03228171	1,81916E-05
	cg13411554	0,40328735	0,08993179	1,85946E-05
	cg05575043	0,36486911	0,08196911	2,11754E-05
	cg14310021	0,34857493	0,07837547	2,14932E-05
BDE-154	cg20136236	0,31566702	0,06451262	3,56173E-06
	cg02466588	0,26914733	0,0585902	1,20492E-05
	cg19720347	0,11978778	0,02629885	1,40678E-05
	cg23706176	-0,299949	0,06594772	1,44406E-05
	cg25683662	0,24050749	0,05325315	1,63923E-05
	cg13062913	0,20758365	0,04596771	1,64213E-05
	cg09285095	0,31591895	0,06996678	1,64595E-05
	cg15134456	-0,357945	0,07938927	1,68905E-05
	cg25780109	-0,2374963	0,05280793	1,76663E-05
	cg06924602	-0,3601048	0,08071305	2,0341E-05
	cg03801924	-0,2962424	0,06648033	2,07811E-05
	cg03940874	-0,3868402	0,08688416	2,10869E-05
	cg07710843	0,13800372	0,03136877	2,59363E-05
	cg10009007	-0,3223394	0,07355106	2,76953E-05
	cg00678890	0,26559988	0,06062325	2,7843E-05
	cg07212852	0,20085233	0,04590962	2,85218E-05
	cg18256856	-0,2737652	0,0627909	3,02295E-05
	cg16712094	0,18233599	0,04197076	3,21087E-05
	cg16684691	-0,2653881	0,06121593	3,32542E-05
	cg22855255	0,23352503	0,05396112	3,42467E-05

Continued on the following page

BDE-183	cg20453042	0,28894084	0,06143405	7,75413E-06
	cg26079864	0,53359048	0,11452357	9,26457E-06
	cg21555123	0,44533854	0,09588594	9,83167E-06
	cg17562250	0,27404726	0,0594459	1,12886E-05
	cg06384026	0,15588329	0,03442006	1,5598E-05
	cg24421265	0,29204936	0,06470844	1,65875E-05
	cg03017264	0,32615244	0,07288842	1,93174E-05
	cg21937462	0,20279726	0,04538176	1,97775E-05
	cg14310021	0,31767976	0,0717454	2,32091E-05
	cg22091565	0,33048967	0,07494362	2,49019E-05
	cg20117519	0,25544695	0,05826218	2,74902E-05
	cg06343669	0,20553723	0,04694638	2,81723E-05
	cg22477463	0,1885924	0,04364593	3,51411E-05
	cg18493250	0,23497298	0,05441117	3,548E-05
	cg16999243	0,37454034	0,08705677	3,77627E-05
	cg03214444	0,26499655	0,06175615	3,9427E-05
	cg08172479	0,34744245	0,08125458	4,17652E-05
	cg25612362	0,27317798	0,06391537	4,20725E-05
	cg24143894	-0,1908371	0,04488941	4,58972E-05
	cg07399928	-0,1851624	0,04368671	4,82007E-05
BDE-209	cg06409164	0,20777551	0,03692273	1,507E-07
	cg16801491	0,15600333	0,03072433	1,64551E-06
	cg22356428	0,33393421	0,06721413	2,60557E-06
	cg09852107	0,17899416	0,03647161	3,35159E-06
	cg21732776	0,1280721	0,02612976	3,4415E-06
	cg00524486	0,177526	0,03660673	4,26443E-06
	cg11702456	0,28783891	0,06000574	5,29566E-06
	cg18404184	0,30980144	0,06601597	8,08901E-06
	cg23806034	0,13056654	0,02823759	1,068E-05
	cg08343644	0,2066366	0,04483999	1,13665E-05
	cg05858126	0,16638051	0,03641219	1,32792E-05
	cg02542953	0,22820754	0,05009261	1,40205E-05
	cg21074797	0,1348953	0,03041447	2,25507E-05
	cg01454153	-0,3409681	0,07732544	2,49333E-05
	cg13421489	0,04602489	0,01052124	2,85747E-05
	cg14142521	0,22692452	0,05190067	2,88183E-05
	cg13347784	0,22293208	0,05101471	2,90797E-05
	cg16658412	0,30796697	0,07062347	3,01431E-05
	cg24443559	-0,4620042	0,10599929	3,03927E-05
	cg06735008	0,16896987	0,038869	3,17625E-05

Continued on the following page

PBDEs	cg15267844	1,31128929	0,23781993	2,49402E-07
	cg01992382	0,46057057	0,08884528	1,04631E-06
	cg27268574	0,19660031	0,03944644	2,43925E-06
	cg16834823	0,41747647	0,08532947	3,57032E-06
	cg21800373	0,19165485	0,03994762	5,27838E-06
	cg15851014	0,26307125	0,05517024	5,95054E-06
	cg14816748	0,12498225	0,02651435	7,43126E-06
	cg01454153	-0,3811482	0,08163465	8,90812E-06
	cg18404184	0,32407271	0,07034682	1,14361E-05
	cg10097464	0,13932414	0,03064174	1,45218E-05
	cg08196740	0,32369745	0,07125401	1,47547E-05
	cg16801491	0,15134915	0,03346368	1,59771E-05
	cg27273140	0,29494422	0,0656216	1,78582E-05
	cg26982927	0,15698867	0,03515689	2,00373E-05
	cg19523085	0,28208358	0,06321382	2,0275E-05
	cg06409164	0,18024596	0,04079722	2,41131E-05
	cg21677976	0,1205776	0,02730967	2,43873E-05
	cg21158631	-0,1985752	0,04501489	2,47588E-05
	cg12058762	-0,2500471	0,05682622	2,58548E-05
	cg04397883	0,1443038	0,03293205	2,77664E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern as fixed effects

Appendix 7. **Top 20 CpGs associated with circulating levels of PBDEs congeners**

	Probes	Coefficients*	SE	P
BDE-28	cg23420164	-0,2480649	0,05271463	7,16025E-06
	cg27128905	0,39274559	0,08465726	9,39769E-06
	cg16461251	0,38116388	0,08351035	1,27293E-05
	cg22151707	-0,1485734	0,03349703	2,12911E-05
	cg07298985	0,81786762	0,18690116	2,69141E-05
	cg04929015	-0,2811104	0,06425435	2,7017E-05
	cg07476726	-0,535364	0,12284623	2,88734E-05
	cg20051358	1,08979167	0,25067525	3,00926E-05
	cg09277673	0,40127965	0,09234823	3,03448E-05
	cg10885779	-0,3204275	0,07382872	3,096E-05
	cg09509943	-0,9432584	0,21737605	3,10633E-05
	cg20966800	0,40698813	0,09403046	3,24291E-05
	cg15892864	0,77520664	0,1791093	3,24466E-05
	cg06085579	-0,2666868	0,06213819	3,73598E-05
	cg20848377	-0,3297637	0,07686684	3,76168E-05
	cg12382431	-0,2709357	0,06371479	4,35222E-05
	cg00475558	0,62239699	0,14645903	4,39754E-05
	cg12818517	-0,2014182	0,04746664	4,50493E-05
	cg12073886	-1,345902	0,31843216	4,8037E-05
	cg16935217	-0,3325964	0,07881674	4,93032E-05
BDE-47	cg25492247	-0,3549654	0,0719646	2,78789E-06
	cg06663935	0,79638576	0,16308674	3,43022E-06
	cg14858675	0,61882096	0,12693491	3,54867E-06
	cg16871475	1,25345544	0,26117573	4,86851E-06
	cg05315595	0,53403063	0,11226077	5,79855E-06
	cg14817226	0,55530163	0,1181225	7,30084E-06
	cg20933239	0,91541783	0,20278887	1,55663E-05
	cg05404233	-0,3139767	0,06987338	1,69035E-05
	cg17834252	1,02249767	0,2296412	1,98913E-05
	cg12664613	1,61948854	0,36592081	2,21174E-05
	cg05646885	-0,322256	0,07335574	2,51636E-05
	cg25270424	0,6722481	0,15473918	3,04508E-05
	cg06335706	-1,4600644	0,33885602	3,49789E-05
	cg07725224	-0,5892405	0,13789287	4,01572E-05
	cg14344448	1,55153413	0,36400354	4,1861E-05
	cg15483273	1,03307519	0,24323478	4,43832E-05
	cg13414654	0,4888586	0,11683446	5,65073E-05
	cg14414338	0,9341759	0,22473219	6,27064E-05
	cg00548060	0,91630623	0,22078582	6,43056E-05
	cg00550498	-0,5286578	0,12740336	6,44823E-05

Continued on the following page

BDE-99	cg17834252	1,64650697	0,33588278	3,16955E-06
	cg26384906	1,10319631	0,22899613	4,51288E-06
	cg14858675	0,84698098	0,18956568	1,871E-05
	cg09961427	1,72459837	0,38958654	2,20345E-05
	cg26893502	0,7930509	0,18042547	2,4928E-05
	cg23916205	-1,12991	0,25731432	2,53514E-05
	cg21549415	0,71708498	0,16396238	2,71754E-05
	cg16327497	-1,2065562	0,27596634	2,73207E-05
	cg00576250	-0,5844777	0,1344277	3,00371E-05
	cg00923230	1,20729663	0,28066411	3,59753E-05
	cg06913229	0,73902334	0,17203868	3,68067E-05
	cg00125706	1,82664363	0,43363405	5,0733E-05
	cg16065213	0,7844213	0,18650565	5,20156E-05
	cg17394189	0,58825976	0,14016873	5,38545E-05
	cg14876453	0,71310607	0,17020217	5,53214E-05
	cg10629020	0,92963355	0,22189281	5,53631E-05
	cg07303829	0,91414105	0,21863986	5,71927E-05
	cg12306307	0,63257923	0,15141244	5,789E-05
	cg14344448	2,26199962	0,54374082	6,19421E-05
	cg06116862	-0,6194192	0,14926485	6,44045E-05
BDE-100	cg02560273	0,93391074	0,21455732	2,94746E-05
	cg26086226	-0,3257132	0,07629806	4,08169E-05
	cg01854076	-0,6541958	0,1537598	4,31316E-05
	cg20933239	0,95013432	0,22457423	4,72736E-05
	cg06587659	0,47492326	0,11245544	4,86751E-05
	cg05295388	-0,3988937	0,09461484	5,00448E-05
	cg18114881	0,41884747	0,09968469	5,28501E-05
	cg27098663	0,84292321	0,20070138	5,32224E-05
	cg16269526	-0,4085236	0,09756023	5,58218E-05
	cg20778915	-1,6178867	0,38655261	5,62438E-05
	cg08254954	-0,412025	0,09871802	5,88005E-05
	cg14817226	0,54852447	0,13168204	6,06717E-05
	cg14414338	1,01894608	0,24711523	7,11816E-05
	cg00906720	0,97251775	0,23643694	7,39588E-05
	cg13569417	-0,3625949	0,08818535	7,43746E-05
	cg25270424	0,70441484	0,17166607	7,6747E-05
	cg02464768	-0,434168	0,10592133	7,80389E-05
	cg02191044	-1,0770415	0,26400757	8,39326E-05
	cg10317119	-1,2804846	0,31455255	8,67383E-05
	cg14167109	-0,3702399	0,09142379	9,38597E-05

Continued on the following page

BDE-153	cg26264999	0,32542382	0,06197149	7,09263E-07
	cg09502865	-0,4949217	0,10118062	3,31333E-06
	cg26189873	-0,5429099	0,11522019	6,98242E-06
	cg20919227	-0,1864994	0,04014459	9,1543E-06
	cg05652609	-0,3459365	0,07521218	1,1043E-05
	cg26421947	-0,2048363	0,04500554	1,34082E-05
	cg05959392	-0,4960297	0,1094352	1,44551E-05
	cg18707191	0,46644004	0,10460164	1,93759E-05
	cg26678852	-0,2889081	0,06558038	2,39698E-05
	cg00370229	-0,5226897	0,119254	2,61786E-05
	cg05088151	-0,2356791	0,05395334	2,77428E-05
	cg01054559	-0,3079039	0,07106326	3,18583E-05
	cg09020104	-0,255701	0,05921644	3,37408E-05
	cg18669948	-0,3103499	0,07280857	4,18397E-05
	cg08282540	-0,2588862	0,06102845	4,52762E-05
	cg06742440	-0,3089802	0,07287375	4,56468E-05
	cg26006682	0,31806093	0,0750475	4,59656E-05
	cg18806716	-0,3637523	0,08630393	5,02737E-05
	cg18834833	0,55411033	0,13156179	5,08537E-05
	cg06496222	-0,2463751	0,05943837	6,55741E-05
BDE-154	cg23619365	-0,4420582	0,08341133	5,73362E-07
	cg00540558	-0,1396153	0,02786476	2,00359E-06
	cg01850798	0,4595655	0,09666969	5,8727E-06
	cg05913250	-0,3341413	0,07286428	1,16705E-05
	cg08733086	-0,2446154	0,05474102	1,86665E-05
	cg01405329	-0,2040126	0,04592366	2,07085E-05
	cg18114881	0,26520525	0,05979697	2,13185E-05
	cg07414863	0,47921031	0,10851527	2,29833E-05
	cg15939915	-0,2443034	0,05661021	3,40756E-05
	cg06476663	0,35203766	0,08160229	3,42709E-05
	cg13576217	0,51645646	0,11989593	3,51512E-05
	cg00007226	-0,1403701	0,03296274	4,25065E-05
	cg26005485	-0,3416935	0,08028976	4,29502E-05
	cg25270424	0,43559195	0,10284165	4,64216E-05
	cg18764771	-0,6448172	0,1522716	4,65839E-05
	cg04290133	0,2938759	0,06955027	4,82732E-05
	cg11688495	0,52754818	0,12502036	4,93356E-05
	cg11028409	-0,2148517	0,05115706	5,32309E-05
	cg11204953	-0,3184451	0,07604012	5,57244E-05
	cg03260790	-0,2653897	0,06360838	5,91399E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI and parity/total breastfeeding duration as fixed effects

Appendix 8. **Top 20 CpGs associated with circulating levels of PBB-153**

Probes	Coefficients*	SE	P
cg23619365	-0,4420582	0,08341133	5,73362E-07
cg00540558	-0,1396153	0,02786476	2,00359E-06
cg01850798	0,4595655	0,09666969	5,8727E-06
cg05913250	-0,3341413	0,07286428	1,16705E-05
cg08733086	-0,2446154	0,05474102	1,86665E-05
cg01405329	-0,2040126	0,04592366	2,07085E-05
cg18114881	0,26520525	0,05979697	2,13185E-05
cg07414863	0,47921031	0,10851527	2,29833E-05
cg15939915	-0,2443034	0,05661021	3,40756E-05
cg06476663	0,35203766	0,08160229	3,42709E-05
cg13576217	0,51645646	0,11989593	3,51512E-05
cg00007226	-0,1403701	0,03296274	4,25065E-05
cg26005485	-0,3416935	0,08028976	4,29502E-05
cg25270424	0,43559195	0,10284165	4,64216E-05
cg18764771	-0,6448172	0,1522716	4,65839E-05
cg04290133	0,2938759	0,06955027	4,82732E-05
cg11688495	0,52754818	0,12502036	4,93356E-05
cg11028409	-0,2148517	0,05115706	5,32309E-05
cg11204953	-0,3184451	0,07604012	5,57244E-05
cg03260790	-0,2653897	0,06360838	5,91399E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI and parity/total breastfeeding duration as fixed effects

Appendix 9. **Top 20 CpGs associated with dietary exposure to PFASs congeners**

	Probes	Coefficients*	SE	P
PFOA	cg08255137	0,20567259	0,03613532	1,1497E-07
	cg10871333	-0,3064394	0,06400266	5,54774E-06
	cg02180424	0,2009751	0,04248686	7,01513E-06
	cg15922246	0,2273075	0,0500071	1,47227E-05
	cg14555350	-0,1658028	0,0370293	1,92473E-05
	cg06715097	0,21920051	0,04932363	2,19472E-05
	cg24389037	-0,3149413	0,07115659	2,35549E-05
	cg03908904	0,22117985	0,0499879	2,36807E-05
	cg10600883	0,18811957	0,04266611	2,51609E-05
	cg27386241	-0,1896615	0,04303511	2,5355E-05
	cg06243540	-0,1465799	0,0334049	2,73152E-05
	cg04857037	0,25487801	0,0587751	3,33334E-05
	cg09366122	-0,3009101	0,06939115	3,33416E-05
	cg04553364	0,2889483	0,06665006	3,34869E-05
	cg00665829	0,18256573	0,04224766	3,53387E-05
	cg03276982	0,30820609	0,07139854	3,59721E-05
	cg03766453	0,29571099	0,06874285	3,81047E-05
	cg17504767	0,21152493	0,04940865	4,1227E-05
	cg24399204	0,22618765	0,05289321	4,19941E-05
	cg09096400	0,52861322	0,12370773	4,25149E-05
PFOS	cg25246012	0,25537909	0,04852336	7,55282E-07
	cg06710082	0,48885907	0,09480325	1,1912E-06
	cg10887021	0,19783827	0,03952889	2,26023E-06
	cg20865068	0,34203764	0,070178	3,89782E-06
	cg24957950	0,35242622	0,07436145	6,76125E-06
	cg19685604	0,72017012	0,15239096	7,14388E-06
	cg08255137	0,20912258	0,0443093	7,32565E-06
	cg21701531	0,2439632	0,05199228	8,18161E-06
	cg27365571	-0,3980988	0,08549246	9,45E-06
	cg23065364	0,32714851	0,07149197	1,30419E-05
	cg07370894	0,22939177	0,05039726	1,43669E-05
	cg08196740	0,35521113	0,07922926	1,88192E-05
	cg08072310	0,2364219	0,05325372	2,23446E-05
	cg03305491	0,40733691	0,09240203	2,52392E-05
	cg02099337	0,43509947	0,09929013	2,7945E-05
	cg14831549	0,34321396	0,07909341	3,29688E-05
	cg03670164	0,28358487	0,06535769	3,30175E-05
	cg08897422	-0,2830722	0,06585254	3,85636E-05
	cg17716817	-0,3965728	0,0923	3,88621E-05
	cg01399598	-0,2476177	0,05786606	4,154E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration, adherence scores to the healthy and Western dietary pattern and lipids as fixed effects

Appendix 10. **Top 20 CpGs associated with circulating levels of PFASs congeners**

	Probes	Estimates	SE	P
PFOA	cg06874740	-0,3702503	0,07266008	1,42218E-06
	cg20828052	0,34523909	0,06963986	2,55848E-06
	cg22860137	0,28511531	0,061776	1,05184E-05
	cg25308242	-0,4746924	0,1040587	1,30504E-05
	cg07025343	0,22795511	0,05137169	2,14153E-05
	cg00115821	-0,3205085	0,07291123	2,52145E-05
	cg10319829	-0,451338	0,10393392	3,10589E-05
	cg21142798	-0,1780092	0,04128497	3,50115E-05
	cg22176913	-0,5007996	0,11629426	3,57526E-05
	cg21499763	0,59313379	0,13777815	3,59362E-05
	cg09386054	-0,2471439	0,05779932	4,02134E-05
	cg00834779	0,31723332	0,07422656	4,05315E-05
	cg10852320	0,17668753	0,0417673	4,7919E-05
	cg21207741	0,15340701	0,03639024	5,06865E-05
	cg14443515	0,19931513	0,04729464	5,09352E-05
	cg07290048	0,63716666	0,15137772	5,19591E-05
	cg15476918	-0,1937254	0,04612828	5,38579E-05
	cg08285915	-0,254341	0,0605676	5,39443E-05
	cg14143723	-0,3606977	0,08602272	5,52422E-05
	cg07775917	-0,2086231	0,04987982	5,75025E-05
PFOS	cg15913831	-0,4015721	0,07712468	8,80964E-07
	cg15507385	0,29495348	0,05676706	9,23702E-07
	cg03202077	0,24260447	0,04689932	1,02021E-06
	cg03158314	-0,2236273	0,04628049	4,32653E-06
	cg22176017	0,17837159	0,03718905	5,01577E-06
	cg02793158	-0,3055703	0,06488142	7,16528E-06
	cg02071825	0,25872771	0,05507853	7,53305E-06
	cg05064673	0,34213082	0,07287844	7,6226E-06
	cg07736327	0,43255907	0,09259885	8,37979E-06
	cg06432204	0,23448868	0,05084579	1,06703E-05
	cg13097573	0,25652278	0,05571483	1,10009E-05
	cg18091163	0,36649984	0,08069343	1,41341E-05
	cg04258138	-0,368994	0,08130744	1,43406E-05
	cg06795069	0,21256984	0,04728207	1,69896E-05
	cg19227131	-0,3260888	0,07287069	1,84596E-05
	cg11279918	0,32182499	0,07282285	2,29975E-05
	cg22369048	0,17628328	0,03994003	2,35088E-05
	cg00187055	-0,2857061	0,06478674	2,38583E-05
	cg03698009	0,34504101	0,07832047	2,42788E-05
	cg11742103	0,56161816	0,12760596	2,46935E-05

*Estimates from linear mixed effect models, one for each congener, with plate and chip as random effects and age, BMI, parity/total breastfeeding duration and lipids as fixed effects

REFERENCES

1. Chatterjee, N. & Walker, G. C. Mechanisms of DNA damage, repair, and mutagenesis: DNA Damage and Repair. *Environ. Mol. Mutagen.* **58**, 235–263 (2017).
2. Cancers_en_France-Essentiel_Faits_et_chiffres-2018.pdf.
3. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
4. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
5. Fouad, Y. A. & Aanei, C. Revisiting the hallmarks of cancer. *Am. J. Cancer Res.* **7**, 1016–1036 (2017).
6. Brücher, B. L. & Jamall, I. S. Epistemology of the origin of cancer: a new paradigm. *BMC Cancer* **14**, 331 (2014).
7. Kennedy, S. R., Loeb, L. A. & Herr, A. J. Somatic mutations in aging, cancer and neurodegeneration. *Mech. Ageing Dev.* **133**, 118–126 (2012).
8. Australian Pancreatic Cancer Genome Initiative *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
9. Tomasetti, C. & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. *Science* **347**, 78–81 (2015).
10. Brücher, B. L. D. M. & Jamall, I. S. Somatic Mutation Theory - Why it's Wrong for Most Cancers. *Cell. Physiol. Biochem.* **38**, 1663–1680 (2016).
11. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering Signatures of Mutational Processes Operative in Human Cancer. *Cell Rep.* **3**, 246–259 (2013).
12. Nik-Zainal, S. *et al.* Mutational Processes Molding the Genomes of 21 Breast Cancers. *Cell* **149**, 979–993 (2012).
13. Fischer, A., Illingworth, C. J., Campbell, P. J. & Mustonen, V. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. *Genome Biol.* **14**, R39 (2013).
14. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
15. Huang, P.-J. *et al.* mSignatureDB: a database for deciphering mutational signatures in human cancers. *Nucleic Acids Res.* **46**, D964–D970 (2018).
16. Mayakonda, A. & Koeffler, H. P. Maftools: Efficient analysis, visualization and summarization of MAF files from large-scale cohort based cancer studies. *bioRxiv* 052662 (2016) doi:10.1101/052662.
17. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, (2017).
18. Kim, S. Y. *et al.* Genomic profiles of a hepatoblastoma from a patient with Beckwith-Wiedemann syndrome with uniparental disomy on chromosome 11p15 and germline mutation of APC and PALB2. *Oncotarget* **8**, (2017).
19. Han, M.-R. *et al.* Mutational signatures and chromosome alteration profiles of squamous cell carcinomas of the vulva. *Exp. Mol. Med.* **50**, e442 (2018).
20. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
21. Alexandrov, L. B. *et al.* The Repertoire of Mutational Signatures in Human Cancer. *bioRxiv* 322859 (2018) doi:10.1101/322859.
22. Zou, X. *et al.* Validating the concept of mutational signatures with isogenic cell models. *Nat. Commun.* **9**, (2018).
23. Perduca, V., Omichessan, H., Baglietto, L. & Severi, G. Mutational and epigenetic signatures in cancer tissue linked to environmental exposures and lifestyle: *Curr. Opin. Oncol.* **30**, 61–67 (2018).
24. Perduca, V. *et al.* Stem cell replication, somatic mutations and role of randomness in the development of cancer. *Eur. J. Epidemiol.* **34**, 439–445 (2019).
25. Kucab, J. E. *et al.* A Compendium of Mutational Signatures of Environmental Agents. *Cell* **177**, 821-836.e16 (2019).
26. Weinhold, B. Epigenetics: The Science of Change. *Environ. Health Perspect.* **114**, (2006).

27. Samaan, R. A. *Dietary fiber for the prevention of cardiovascular disease: fiber's interaction between gut micoflora, sugar metabolism, weight control and cardiovascular health*. (2017).
28. Handy, D. E., Castro, R. & Loscalzo, J. Epigenetic Modifications: Basic Mechanisms and Role in Cardiovascular Disease. *Circulation* **123**, 2145–2156 (2011).
29. Aguilera, O., Fernández, A. F., Muñoz, A. & Fraga, M. F. Epigenetics and environment: a complex relationship. *J. Appl. Physiol.* **109**, 243–251 (2010).
30. Hamilton, J. P. Epigenetics: Principles and Practice. *Dig. Dis.* **29**, 130–135 (2011).
31. Mahmoud, A. & Ali, M. Methyl Donor Micronutrients that Modify DNA Methylation and Cancer Outcome. *Nutrients* **11**, 608 (2019).
32. Wajed, S. A., Laird, P. W. & DeMeester, T. R. DNA methylation: an alternative pathway to cancer. *Ann. Surg.* **234**, 10–20 (2001).
33. Sharp, A. J. *et al.* DNA methylation profiles of human active and inactive X chromosomes. *Genome Res.* **21**, 1592–1600 (2011).
34. Robertson, K. D. DNA methylation and human disease. *Nat. Rev. Genet.* **6**, 597–610 (2005).
35. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
36. Vijg, J. & Suh, Y. Genome Instability and Aging. *Annu. Rev. Physiol.* **75**, 645–668 (2013).
37. Victorelli, S. & Passos, J. F. Telomeres and Cell Senescence - Size Matters Not. *EBioMedicine* **21**, 14–20 (2017).
38. Yeh, J.-K. & Wang, C.-Y. Telomeres and Telomerase in Cardiovascular Diseases. *Genes* **7**, 58 (2016).
39. Ishida, T., Ishida, M., Tashiro, S., Yoshizumi, M. & Kihara, Y. Role of DNA damage in cardiovascular disease. *Circ. J. Off. J. Jpn. Circ. Soc.* **78**, 42–50 (2014).
40. Barzilai, A., Schumacher, B. & Shiloh, Y. Genome instability: Linking ageing and brain degeneration. *Mech. Ageing Dev.* **161**, 4–18 (2017).
41. Hou, Y., Song, H., Croteau, D. L., Akbari, M. & Bohr, V. A. Genome instability in Alzheimer disease. *Mech. Ageing Dev.* **161**, 83–94 (2017).
42. Esteller, M. & Almouzni, G. How epigenetics integrates nuclear functions. Workshop on epigenetics and chromatin: transcriptional regulation and beyond. *EMBO Rep.* **6**, 624–628 (2005).
43. You, J. S. & Jones, P. A. Cancer Genetics and Epigenetics: Two Sides of the Same Coin? *Cancer Cell* **22**, 9–20 (2012).
44. Sandoval, J. & Esteller, M. Cancer epigenomics: beyond genomics. *Curr. Opin. Genet. Dev.* **22**, 50–55 (2012).
45. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome — biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
46. Barros-Silva, D., Marques, C., Henrique, R. & Jerónimo, C. Profiling DNA Methylation Based on Next-Generation Sequencing Approaches: New Insights and Clinical Applications. *Genes* **9**, 429 (2018).
47. Fan, S. & Chi, W. Methods for genome-wide DNA methylation analysis in human cancer. *Brief. Funct. Genomics* elw010 (2016) doi:10.1093/bfgp/elw010.
48. Zafon, C., Gil, J., Pérez-González, B. & Jordà, M. DNA methylation in thyroid cancer. *Endocr. Relat. Cancer* **26**, R415–R439 (2019).
49. Tiffon, C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. *Int. J. Mol. Sci.* **19**, 3425 (2018).
50. Mahna, D., Puri, S. & Sharma, S. DNA methylation signatures: Biomarkers of drug and alcohol abuse. *Mutat. Res. Mutat. Res.* **777**, 19–28 (2018).
51. Gensous, N. *et al.* The Impact of Caloric Restriction on the Epigenetic Signatures of Aging. *Int. J. Mol. Sci.* **20**, 2022 (2019).
52. Jin, F. *et al.* Tobacco-Specific Carcinogens Induce Hypermethylation, DNA Adducts, and DNA Damage in Bladder Cancer. *Cancer Prev. Res. (Phila. Pa.)* **10**, 588–597 (2017).
53. Baglietto, L. *et al.* DNA methylation changes measured in pre-diagnostic peripheral blood samples are associated with smoking and lung cancer risk: DNA methylation changes in pre-diagnostic blood samples and lung cancer risk. *Int. J. Cancer* **140**, 50–61 (2017).
54. Fasanelli, F. *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).

55. Xu, X. *et al.* High intakes of choline and betaine reduce breast cancer mortality in a population-based study. *FASEB J.* **23**, 4022–4028 (2009).
56. Zhou, R. *et al.* Higher dietary intakes of choline and betaine are associated with a lower risk of primary liver cancer: a case-control study. *Sci. Rep.* **7**, 679 (2017).
57. Serra-Majem, L. *et al.* Benefits of the Mediterranean diet: Epidemiological and molecular aspects. *Mol. Aspects Med.* **67**, 1–55 (2019).
58. Arpón, A. *et al.* Adherence to Mediterranean diet is associated with methylation changes in inflammation-related genes in peripheral blood cells. *J. Physiol. Biochem.* **73**, 445–455 (2016).
59. Kortenkamp, A. *et al.* International Programme on Chemical Safety (WHO/UNEP/ILO) "Global assessment of the state-of-the-science of endocrine disruptors". IPCS, 2002.
60. *Brominated Flame Retardants*. vol. 16 (Springer Berlin Heidelberg, 2011).
61. Jones, K. C. & de Voogt, P. Persistent organic pollutants (POPs): state of the science. *Environ. Pollut. Barking Essex 1987* **100**, 209–221 (1999).
62. Geyer, H. J. *et al.* Terminal elimination half-lives of the brominated flame retardants TBBPA, HBCD, and lower brominated PBDEs in humans. *ORGANOHALOGEN Compd.* **66**, 7 (2004).
63. Malliari, E. & Kalantzi, O.-I. Children's exposure to brominated flame retardants in indoor environments - A review. *Environ. Int.* **108**, 146–169 (2017).
64. Eriksson, P., Jakobsson, E. & Fredriksson, A. Brominated flame retardants: a novel class of developmental neurotoxicants in our environment? *Environ. Health Perspect.* **109**, 903–908 (2001).
65. Eriksson, P., Viberg, H., Jakobsson, E., Orn, U. & Fredriksson, A. A Brominated Flame Retardant, 2,2`,4,4`,5-Pentabromodiphenyl Ether: Uptake, Retention, and Induction of Neurobehavioral Alterations in Mice during a Critical Phase of Neonatal Brain Development. *Toxicol. Sci.* **67**, 98–103 (2002).
66. He, J. *et al.* Chronic zebrafish low dose decabrominated diphenyl ether (BDE-209) exposure affected parental gonad development and locomotion in F1 offspring. *Ecotoxicology* **20**, 1813–1822 (2011).
67. Hendriks, H. S. & Westerink, R. H. S. Neurotoxicity and risk assessment of brominated and alternative flame retardants. *Neurotoxicol. Teratol.* **52**, 248–269 (2015).
68. Khalil, A. *et al.* Perinatal exposure to 2,2′,4′4′ −Tetrabromodiphenyl ether induces testicular toxicity in adult rats. *Toxicology* **389**, 21–30 (2017).
69. WHO (2009) Biomonitoring of human milk for persistent organic pollutants (POPs).
70. Müller, M. H. B. *et al.* Brominated flame retardants (BFRs) in breast milk and associated health risks to nursing infants in Northern Tanzania. *Environ. Int.* **89–90**, 38–47 (2016).
71. Chen, T., Huang, M., Li, J., Li, J. & Shi, Z. Polybrominated diphenyl ethers and novel brominated flame retardants in human milk from the general population in Beijing, China: Occurrence, temporal trends, nursing infants' exposure and risk assessment. *Sci. Total Environ.* **689**, 278–286 (2019).
72. Leonetti, C. *et al.* Brominated flame retardants in placental tissues: associations with infant sex and thyroid hormone endpoints. *Environ. Health* **15**, 113 (2016).
73. Liu, S. *et al.* Association of polybrominated diphenylethers (PBDEs) and hydroxylated metabolites (OH-PBDEs) serum levels with thyroid function in thyroid cancer patients. *Environ. Res.* **159**, 1–8 (2017).
74. Byrne, S. C. *et al.* Associations between serum polybrominated diphenyl ethers and thyroid hormones in a cross sectional study of a remote Alaska Native population. *Sci. Rep.* **8**, 2198 (2018).
75. Shrestha, S. *et al.* Perfluoroalkyl substances and thyroid function in older adults. *Environ. Int.* **75**, 206–214 (2015).
76. Ongono, J. S. *et al.* Dietary exposure to brominated flame retardants and risk of type 2 diabetes in the French E3N cohort. *Environ. Int.* **123**, 54–60 (2019).
77. Wöhrnschimmel, H. *et al.* Ten years after entry into force of the Stockholm Convention: What do air monitoring data tell about its effectiveness? *Environ. Pollut.* **217**, 149–158 (2016).
78. Guo, W. *et al.* PBDE levels in breast milk are decreasing in California. *Chemosphere* **150**, 505–513 (2016).
79. Zhang, L. *et al.* Polybrominated diphenyl ethers and indicator polychlorinated biphenyls in human milk from China under the Stockholm Convention. *Chemosphere* **189**, 32–38 (2017).
80. Risk to human health related to the presence of perfluorooctane sulfonic acid and perfluorooctanoic acid in food 2018 - EFSA Journal.

81. Liu, W., Wu, J., He, W. & Xu, F. A review on perfluoroalkyl acids studies: Environmental behaviors, toxic effects, and ecological and health risks. *Ecosyst. Health Sustain.* **5**, 1–19 (2019).
82. Bach, C. *et al.* The impact of two fluoropolymer manufacturing facilities on downstream contamination of a river and drinking water resources with per- and polyfluoroalkyl substances. *Environ. Sci. Pollut. Res.* **24**, 4916–4925 (2017).
83. Sznajder-Katarzyńska, K., Surma, M. & Cieślik, I. A Review of Perfluoroalkyl Acids (PFAAs) in terms of Sources, Applications, Human Exposure, Dietary Intake, Toxicity, Legal Regulation, and Methods of Determination. *J. Chem.* **2019**, 1–20 (2019).
84. Tanner, E. M. *et al.* Occupational exposure to perfluoroalkyl substances and serum levels of perfluorooctanesulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) in an aging population from upstate New York: a retrospective cohort study. *Int. Arch. Occup. Environ. Health* **91**, 145–154 (2018).
85. Alexander, B. H. & Olsen, G. W. Bladder Cancer in Perfluorooctanesulfonyl Fluoride Manufacturing Workers. *Ann. Epidemiol.* **17**, 471–478 (2007).
86. Braun, J. M. Early-life exposure to EDCs: role in childhood obesity and neurodevelopment. *Nat. Rev. Endocrinol.* **13**, 161–173 (2017).
87. Bach, C. C. *et al.* Perfluoroalkyl and polyfluoroalkyl substances and human fetal growth: A systematic review. *Crit. Rev. Toxicol.* **45**, 53–67 (2015).
88. Jian, J.-M. *et al.* A short review on human exposure to and tissue distribution of per- and polyfluoroalkyl substances (PFASs). *Sci. Total Environ.* **636**, 1058–1069 (2018).
89. Olsen, G. W. *et al.* Half-Life of Serum Elimination of Perfluorooctanesulfonate,Perfluorohexanesulfonate, and Perfluorooctanoate in Retired Fluorochemical Production Workers. *Environ. Health Perspect.* **115**, 1298–1305 (2007).
90. Li, Y. *et al.* Half-lives of PFOS, PFHxS and PFOA after end of exposure to contaminated drinking water. *Occup. Environ. Med.* **75**, 46–51 (2018).
91. Lau, C. *et al.* Perfluoroalkyl Acids: A Review of Monitoring and Toxicological Findings. *Toxicol. Sci.* **99**, 366–394 (2007).
92. Chen, J. *et al.* Early life perfluorooctanesulphonic acid (PFOS) exposure impairs zebrafish organogenesis. *Aquat. Toxicol.* **150**, 124–132 (2014).
93. Chen, J. *et al.* Chronic perfluorooctanesulphonic acid (PFOS) exposure produces estrogenic effects in zebrafish. *Environ. Pollut.* **218**, 702–708 (2016).
94. Mancini, F. R. *et al.* Perfluorinated alkylated substances serum concentration and breast cancer risk: Evidence from a nested case-control study in the French E3N cohort. *Int. J. Cancer* ijc.32357 (2019) doi:10.1002/ijc.32357.
95. Wielsøe, M., Kern, P. & Bonefeld-Jørgensen, E. C. Serum levels of environmental pollutants is a risk factor for breast cancer in Inuit: a case control study. *Environ. Health* **16**, 56 (2017).
96. Hurley, S. *et al.* Breast cancer risk and serum levels of per- and poly-fluoroalkyl substances: a case-control study nested in the California Teachers Study. *Environ. Health* **17**, 83 (2018).
97. Cardenas, A. *et al.* Associations of Perfluoroalkyl and Polyfluoroalkyl Substances With Incident Diabetes and Microvascular Disease. *Diabetes Care* **42**, 1824–1832 (2019).
98. Christensen, K. Y., Raymond, M., Thompson, B. A. & Anderson, H. A. Perfluoroalkyl substances in older male anglers in Wisconsin. *Environ. Int.* **91**, 312–318 (2016).
99. Sun, Q. *et al.* Plasma Concentrations of Perfluoroalkyl Substances and Risk of Type 2 Diabetes: A Prospective Investigation among U.S. Women. *Environ. Health Perspect.* **126**, 037001 (2018).
100. Donat-Vargas, C. *et al.* Perfluoroalkyl substances and risk of type II diabetes: A prospective nested case-control study. *Environ. Int.* **123**, 390–398 (2019).
101. Karnes, C., Winquist, A. & Steenland, K. Incidence of type II diabetes in a cohort with substantial exposure to perfluorooctanoic acid. *Environ. Res.* **128**, 78–83 (2014).
102. Alvarado-Cruz, I., Alegría-Torres, J. A., Montes-Castro, N., Jiménez-Garza, O. & Quintanilla-Vega, B. Environmental Epigenetic Changes, as Risk Factors for the Development of Diseases in Children: A Systematic Review. *Ann. Glob. Health* **84**, 212–224 (2018).
103. Deininger, P. Alu elements: know the SINEs. *Genome Biol.* **12**, 236 (2011).
104. Huen, K. *et al.* Effects of age, sex, and persistent organic pollutants on DNA methylation in children: DNA Methylation in Children. *Environ. Mol. Mutagen.* **55**, 209–222 (2014).
105. Kim, K.-Y. *et al.* Association of Low-Dose Exposure to Persistent Organic Pollutants with Global DNA Hypomethylation in Healthy Koreans. *Environ. Health Perspect.* **118**, 370–374 (2010).

106. Liu, C.-Y., Chen, P.-C., Lien, P.-C. & Liao, Y.-P. Prenatal Perfluorooctyl Sulfonate Exposure and Alu DNA Hypomethylation in Cord Blood. *Int. J. Environ. Res. Public. Health* **15**, 1066 (2018).
107. Guerrero-Preston, R. *et al.* Global DNA hypomethylation is associated with in utero exposure to cotinine and perfluorinated alkyl compounds. *Epigenetics* **5**, 539–546 (2010).
108. Karimi, M., Luttropp, K. & Ekström, T. J. Global DNA Methylation Analysis Using the Luminometric Methylation Assay. in *Epigenetics Protocols* (ed. Tollefsbol, T. O.) vol. 791 135–144 (Humana Press, 2011).
109. Dao, T., Hong, X., Wang, X. & Tang, W.-Y. Aberrant 5'-CpG Methylation of Cord Blood TNFα Associated with Maternal Exposure to Polybrominated Diphenyl Ethers. *PLOS ONE* **10**, e0138815 (2015).
110. Consales, C. *et al.* Exposure to persistent organic pollutants and sperm DNA methylation changes in Arctic and European populations: Sperm DNA Methylation and POP Exposure. *Environ. Mol. Mutagen.* **57**, 200–209 (2016).
111. Leter, G. *et al.* Exposure to Perfluoroalkyl Substances and Sperm DNA Global Methylation in Arctic and European Populations: Sperm DNA Methylation and PFASs Exposure. *Environ. Mol. Mutagen.* **55**, 591–600 (2014).
112. Soubry, A. *et al.* Human exposure to flame-retardants is associated with aberrant DNA methylation at imprinted genes in sperm. *Environ. Epigenetics* **3**, (2017).
113. Perduca, V., Omichessan, H., Baglietto, L. & Severi, G. Mutational and epigenetic signatures in cancer tissue linked to environmental exposures and lifestyle: *Curr. Opin. Oncol.* **30**, 61–67 (2018).
114. Vogelstein, B. & Kinzler, K. W. Carcinogens leave fingerprints. *Nature* **355**, 209–210 (1992).
115. Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Curr. Opin. Genet. Dev.* **24**, 52–60 (2014).
116. Nik-Zainal, S. & Morganella, S. Mutational Signatures in Breast Cancer: The Problem at the DNA Level. *Clin. Cancer Res.* **23**, 2617–2629 (2017).
117. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
118. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science* **354**, 618–622 (2016).
119. Denissenko, M. F., Pao, A., Tang, M. -s. & Pfeifer, G. P. Preferential Formation of Benzo[a]pyrene Adducts at Lung Cancer Mutational Hotspots in P53. *Science* **274**, 430–432 (1996).
120. Huang, M. N. *et al.* Genome-scale mutational signatures of aflatoxin in cells, mice, and human tumors. *Genome Res.* **27**, 1475–1486 (2017).
121. ICGC Prostate Group *et al.* Mutational signatures of ionizing radiation in second malignancies. *Nat. Commun.* **7**, 12605 (2016).
122. Davidson, P. R., Sherborne, A. L., Taylor, B., Nakamura, A. O. & Nakamura, J. L. A pooled mutational analysis identifies ionizing radiation-associated mutational signatures conserved between mouse and human malignancies. *Sci. Rep.* **7**, 7645 (2017).
123. Castells, X. *et al.* Low-Coverage Exome Sequencing Screen in Formalin-Fixed Paraffin-Embedded Tumors Reveals Evidence of Exposure to Carcinogenic Aristolochic Acid. *Cancer Epidemiol. Biomarkers Prev.* **24**, 1873–1881 (2015).
124. Poon, S. L. *et al.* Genome-Wide Mutational Signatures of Aristolochic Acid and Its Application as a Screening Tool. *Sci. Transl. Med.* **5**, 197ra101-197ra101 (2013).
125. Hoang, M. L. *et al.* Mutational Signature of Aristolochic Acid Exposure as Revealed by Whole-Exome Sequencing. *Sci. Transl. Med.* **5**, 197ra102-197ra102 (2013).
126. Poon, S. L. *et al.* Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med.* **7**, (2015).
127. Hoang, M. L. *et al.* Aristolochic Acid in the Etiology of Renal Cell Carcinoma. *Cancer Epidemiol. Biomarkers Prev.* **25**, 1600–1608 (2016).
128. Turesky, R. J. *et al.* Aristolochic acid exposure in Romania and implications for renal cell carcinoma. *Br. J. Cancer* **114**, 76–80 (2016).
129. Scelo, G. *et al.* Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.* **5**, 5135 (2014).
130. Barrow, T. M. & Michels, K. B. Epigenetic epidemiology of cancer. *Biochem. Biophys. Res. Commun.* **455**, 70–83 (2014).

131. Timp, W. *et al.* Large hypomethylated blocks as a universal defining epigenetic alteration in human solid tumors. *Genome Med.* **6**, 61 (2014).
132. Kurdyukov, S. & Bullock, M. DNA Methylation Analysis: Choosing the Right Method. *Biology* **5**, 3 (2016).
133. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **9**, 436–447 (2016).
134. Guida, F. *et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24**, 2349–2359 (2015).
135. Teschendorff, A. E. *et al.* Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer. *JAMA Oncol.* **1**, 476 (2015).
136. Stueve, T. R. *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.* **26**, 3014–3027 (2017).
137. Battram, T. *et al.* Appraising the causal relevance of DNA methylation for risk of lung cancer. *Int. J. Epidemiol.* **48**, 1493–1504 (2019).
138. Secretan, B. *et al.* A review of human carcinogens—Part E: tobacco, areca nut, alcohol, coal smoke, and salted fish. *Lancet Oncol.* **10**, 1033–1034 (2009).
139. Les cancers attribuables au mode de vie et à l'environnement en France métropolitaine.pdf.
140. Tomasetti, C. & Vogelstein, B. Cancer risk: Role of environment--Response. *Science* **347**, 729–731 (2015).
141. Tomasetti, C. & Vogelstein, B. Musings on the theory that variation in cancer risk among tissues can be explained by the number of divisions of normal stem cells. *ArXiv150105035 Q-Bio Stat* (2015).
142. Tomasetti, C. & Vogelstein, B. On the slope of the regression between stem cell divisions and cancer risk, and the lack of correlation between stem cell divisions and environmental factors-associated cancer risk. *PLOS ONE* **12**, e0175535 (2017).
143. Tomasetti, C., Li, L. & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. *Science* **355**, 1330–1334 (2017).
144. Weinberg, C. R. & Zaykin, D. Is Bad Luck the Main Cause of Cancer? *JNCI J. Natl. Cancer Inst.* **107**, djv125–djv125 (2015).
145. Kelly-Irving, M., Delpierre, C. & Vineis, P. Beyond bad luck: induced mutations and hallmarks of cancer. *Lancet Oncol.* **18**, 999–1000 (2017).
146. López-Lázaro, M. The stem cell division theory of cancer. *Crit. Rev. Oncol. Hematol.* **123**, 95–113 (2018).
147. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2016).
148. Little, M. P., Hendry, J. H. & Puskin, J. S. Lack of Correlation between Stem-Cell Proliferation and Radiation- or Smoking-Associated Cancer Risk. *PLOS ONE* **11**, e0150335 (2016).
149. Agudo, A. *et al.* Impact of Cigarette Smoking on Cancer Risk in the European Prospective Investigation into Cancer and Nutrition Study. *J. Clin. Oncol.* **30**, 4550–4557 (2012).
150. Doll, R., Peto, R., Boreham, J. & Sutherland, I. Mortality in relation to smoking: 50 years' observations on male British doctors. *BMJ* **328**, 1519 (2004).
151. Rosales, R. A., Drummond, R. D., Valieris, R., Dias-Neto, E. & da Silva, I. T. signeR: an empirical Bayesian approach to mutational signature discovery. *Bioinformatics* **33**, 8–16 (2017).
152. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants: Fig. 1. *Bioinformatics* btv408 (2015) doi:10.1093/bioinformatics/btv408.
153. Shiraishi, Y., Tremmel, G., Miyano, S. & Stephens, M. A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS Genet.* **11**, e1005657 (2015).
154. Kim, J. *et al.* Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
155. Kasar, S. *et al.* Whole-genome sequencing reveals activation-induced cytidine deaminase signatures during indolent chronic lymphocytic leukaemia evolution. *Nat. Commun.* **6**, (2015).
156. Tan, V. Y. F. & Fevotte, C. Automatic Relevance Determination in Nonnegative Matrix Factorization with the /spl beta/-Divergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1592–1605 (2013).

157. Fantini, D., Huang, S., Asara, J. M., Bagchi, S. & Raychaudhuri, P. Chromatin association of XRCC5/6 in the absence of DNA damage depends on the XPE gene product DDB2. *Mol. Biol. Cell* **28**, 192–200 (2017).
158. Carlson, J., Li, J. Z. & Zöllner, S. Helmsman: fast and efficient mutation signature analysis for massive sequencing datasets. *BMC Genomics* **19**, (2018).
159. Ramazzotti, D., Lal, A., Liu, K., Tibshirani, R. & Sidow, A. De Novo Mutational Signature Discovery in Tumor Genomes using SparseSignatures. *bioRxiv* (2018) doi:10.1101/384834.
160. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, (2016).
161. Lynch, A. G. Decomposition of mutational context signatures using quadratic programming methods. *F1000Research* **5**, 1253 (2016).
162. Huang, X., Wojtowicz, D. & Przytycka, T. M. Detecting presence of mutational signatures in cancer with confidence. *Bioinformatics* **34**, 330–337 (2018).
163. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med.* **10**, (2018).
164. Huebschmann, D., Gu, Z. & Schelsner, M. YAPSA: Yet Another Package for Signature Analysis R package version 1.6.0. (2015).
165. Krüger, S. & Piro, R. M. Identification of mutational signatures active in individual tumors. doi:10.7287/peerj.preprints.3257v1.
166. Gori, K. & Baez-Ortega, A. sigfit: flexible Bayesian inference of mutational signatures. *bioRxiv* (2018) doi:10.1101/372896.
167. Ardin, M. *et al.* MutSpec: a Galaxy toolbox for streamlined analyses of somatic mutation spectra in human and mouse cancer genomes. *BMC Bioinformatics* **17**, (2016).
168. Goncearenco, A. *et al.* Exploring background mutational processes to decipher cancer genetic heterogeneity. *Nucleic Acids Res.* **45**, W514–W522 (2017).
169. Lee, J. *et al.* Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. *Nucleic Acids Res.* **46**, W102–W108 (2018).
170. Díaz-Gay, M. *et al.* Mutational Signatures in Cancer (MuSiCa): a web application to implement mutational signatures analysis in cancer samples. *BMC Bioinformatics* **19**, (2018).
171. Fan, Y. *et al.* MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data. *Genome Biol.* **17**, (2016).
172. Liao, X. & Meyer, M. C. **coneproj** : An *R* Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. *J. Stat. Softw.* **61**, (2014).
173. Baez-Ortega, A. & Gori, K. Computational approaches for discovery of mutational signatures in cancer. *Brief. Bioinform.* (2017) doi:10.1093/bib/bbx082.
174. Antignac, J.-P. *et al.* Exposure assessment of French women and their newborn to brominated flame retardants: Determination of tri- to deca- polybromodiphenylethers (PBDE) in maternal adipose tissue, serum, breast milk and cord serum. *Environ. Pollut.* **157**, 164–173 (2009).
175. Cariou, R. *et al.* Perfluoroalkyl acid (PFAA) levels and profiles in breast milk, maternal and cord serum of French women and their newborns. *Environ. Int.* **84**, 71–81 (2015).
176. Sirot, V. *et al.* Core food of the French food supply: second Total Diet Study. *Food Addit. Contam. Part A* **26**, 623–639 (2009).
177. van Liere, M. Relative validity and reproducibility of a French dietary history questionnaire. *Int. J. Epidemiol.* **26**, 128S – 136 (1997).
178. Cariou, R. *et al.* New multiresidue analytical method dedicated to trace level measurement of brominated flame retardants in human biological matrices. *J. Chromatogr. A* **1100**, 144–152 (2005).
179. Akins, J. R., Waldrep, K. & Bernert, J. T. The estimation of total serum lipids by a completely enzymatic 'summation' method. *Clin. Chim. Acta* **184**, 219–226 (1989).
180. Fasanelli, F. *et al.* Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nat. Commun.* **6**, 10192 (2015).
181. Fasanelli, F. *et al.* DNA methylation, colon cancer and Mediterranean diet: results from the EPIC-Italy cohort. *Epigenetics* **14**, 977–988 (2019).
182. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using **lme4**. *J. Stat. Softw.* **67**, (2015).

183. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
184. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).
185. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
186. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
187. Edefonti, V. *et al.* Nutrient dietary patterns and the risk of breast and ovarian cancers. *Int. J. Cancer* **122**, 609–613 (2008).
188. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
189. Kesse, E., Clavel-Chapelon, F. & Boutron-Ruault, M. Dietary Patterns and Risk of Colorectal Tumors: A Cohort of French Women of the National Education System (E3N). *Am. J. Epidemiol.* **164**, 1085–1093 (2006).
190. Jain, R. B. & Ducatman, A. Roles of gender and obesity in defining correlations between perfluoroalkyl substances and lipid/lipoproteins. *Sci. Total Environ.* **653**, 74–81 (2019).
191. Bassler, J. *et al.* Environmental perfluoroalkyl acid exposures are associated with liver disease characterized by apoptosis and altered serum adipocytokines. *Environ. Pollut.* **247**, 1055–1063 (2019).
192. Thompson, P. A. *et al.* Environmental immune disruptors, inflammation and cancer risk. *Carcinogenesis* **36**, S232–S253 (2015).
193. Zota, A. R. *et al.* Association between persistent endocrine-disrupting chemicals (PBDEs, OH-PBDEs, PCBs, and PFASs) and biomarkers of inflammation and cellular aging during pregnancy and postpartum. *Environ. Int.* **115**, 9–20 (2018).
194. Linares, V., Bellés, M. & Domingo, J. L. Human exposure to PBDE and critical evaluation of health hazards. *Arch. Toxicol.* **89**, 335–356 (2015).
195. Sarkar, D., Joshi, D. & Singh, S. K. Maternal BDE-209 exposure during lactation causes testicular and epididymal toxicity through increased oxidative stress in peripubertal mice offspring. *Toxicol. Lett.* **311**, 66–79 (2019).
196. Wan, H. T. *et al.* PFOS-induced hepatic steatosis, the mechanistic actions on β-oxidation and lipid transport. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1820**, 1092–1101 (2012).
197. Lv, Z. *et al.* Glucose and lipid homeostasis in adult rat is impaired by early-life exposure to perfluorooctane sulfonate: Early-Life PFOS Exposure Increase the Risk of Insulin Resistance in Adulthood. *Environ. Toxicol.* **28**, 532–542 (2013).
198. Liang, X., Xie, G., Wu, X., Su, M. & Yang, B. Effect of prenatal PFOS exposure on liver cell function in neonatal mice. *Environ. Sci. Pollut. Res.* **26**, 18240–18246 (2019).
199. VanderWeele, T. J. & Vansteelandt, S. Odds Ratios for Mediation Analysis for a Dichotomous Outcome. *Am. J. Epidemiol.* **172**, 1339–1348 (2010).
200. Salihovic, S. *et al.* Perfluoroalkyl substances (PFAS) including structural PFOS isomers in plasma from elderly men and women from Sweden: Results from the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS). *Environ. Int.* **82**, 21–27 (2015).
201. Kim, S.-J., Choi, E.-J., Choi, G.-W., Lee, Y.-B. & Cho, H.-Y. Exploring sex differences in human health risk assessment for PFNA and PFDA using a PBPK model. *Arch. Toxicol.* **93**, 311–330 (2019).

Title: Applications of genomic and epigenomic signatures to identify markers of exogenous exposures and elucidate their potential role in cancer aetiology.

Context and aim: Several risks factors have been identified for cancer, and it has been estimated that more than 40% of cases in developed countries are preventable through the modulation of known modifiable risk factors. The overall objective of this thesis was to demonstrate that the analysis of genomic and epigenomic data integrated with well-characterised exposure and lifestyle data may be used to identify markers of environmental exposures and lifestyle and may contribute to increase our understanding of cancer aetiology.
Results: We first describe how genomic and epigenomic signatures can be used to identify markers of exposure and decipher the aetiology of cancer. Then, we adopt the mutational signatures framework to contribute to the debate about the "bad luck" hypothesis for cancer and demonstrate that tobacco-related mutations are more strongly correlated with cancer risk than random mutations. We introduce a probabilistic model for the simulation of mutational signature data and compare the performance of the available methods for the identification of mutational signatures using both simulated and real data. Additionally, we introduce a new method for the identification of such signatures. Finally, we use methylation array data in an epidemiological study within the E3N cohort to investigate the association between exposure to Brominated Flame Retardants and Per- and polyfluoroalkyl substances, two organic pollutants that are known endocrine disrupting chemicals, and methylation in DNA from blood. Overall, our study does not provide evidence of methylation alterations at the level of the whole genome, in regions or in single CpGs. Suggestive evidence of alterations in the methylation of genes within plausible biological pathways (e.g. androgen response) warrants further investigations.
Conclusion: Our work on the methodological aspects of mutational signature research introduces an original framework for measuring the performance of tools for the identification of mutational signatures that may serve as reference for future methodological or applied research. Our applications of both mutational signature and methylome research demonstrate the usefulness of such tools to assess exposures and elucidate their role in cancer aetiology.

Keywords **:** mutational signatures, DNA methylation, endocrine disruptors, epidemiology, lifestyle

Titre : Utilisation des signatures génomiques et épigenomiques dans le but d'identifier des marqueurs d'expositions exogènes et d'évaluer leur rôle dans l'étiologie du cancer.

Contexte et objectif : Plusieurs facteurs de risque de cancer ont été identifiés et il a été estimé que plus de 40% des cas dans les pays développés pourraient être évités en modifiant les facteurs de risque connus. L'objectif général de cette thèse était de démontrer que l'intégration de données génomiques et épigénomiques aux données détaillées sur les expositions environnementales et le mode de vie peut être utile pour identifier des biomarqueurs de ces facteurs et contribuer à augmenter notre connaissance de l'étiologie du cancer.
Résultats : Dans un premier temps, nous décrivons comment les signatures génomiques et épigénomiques peuvent être utilisées pour identifier des marqueurs d'exposition et déchiffrer l'étiologie du cancer. Ensuite, nous contribuons au débat relatif à l'hypothèse de la chance dans le développement du cancer et démontrons que les mutations induites par le tabagisme sont plus prédictives du risque de cancer que les mutations aléatoires. Nous introduisons un modèle probabiliste pour la simulation de données mutationnelles et comparons la performance des outils d'identification de ces signatures avec des données réelles et simulées. De plus, nous introduisons une nouvelle méthode pour l'identification des signatures mutationnelles. Enfin, nous utilisons les données de méthylation de la cohorte E3N pour étudier le lien entre l'exposition aux retardateurs de flamme bromés et aux composés perfluorés, deux substances classées parmi les perturbateurs endocriniens, et la méthylation de l'ADN sanguin. Globalement, notre étude ne fournit aucune preuve d'altérations globales du méthylome ou d'altérations à l'échelle des CpGs. Cependant, certains résultats suggèrent l'existence d'altérations de la méthylation de gènes impliqués dans des voies biologiques (ex., la réponse aux androgènes) et nécessitent des recherches supplémentaires**.**
Conclusion : Ce travail contribue à la recherche méthodologique portant sur les signatures mutationnelles en introduisant un protocole de mesure de performance et d'identification des signatures mutationnelles pouvant servir de référence à de futures études méthodologiques ou appliquées. Nos recherches sur les signatures mutationnelles et le méthylome démontrent l'utilité de tels outils pour évaluer les expositions et élucider leur rôle dans l'étiologie du cancer.

Mots clés **:** signatures mutationnelles, méthylation de l'ADN, perturbateurs endocriniens, épidémiologie, mode de vie