Analyse de la médiation des effets d'expositions environnementales sur la santé via la méthylation de l'ADN : Application à l'exposition prénatale au tabagisme et à la pollution atmosphérique et la santé de l'enfant

Basile Jumentier

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : ISCE - Ingénierie pour la Santé la Cognition et l'Environnement
Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Unité de recherche : Translational Innovation in Medicine and Complexity

Analyse de la médiation des effets d'expositions environnementales sur la santé via la méthylation de l'ADN : Application à l'exposition prénatale au tabagisme et à la pollution atmosphérique et la santé de l'enfant

Analysis of the mediation of effects of environmental exposures on health via DNA methylation: Application to prenatal exposure to tobacco and air pollution and child health

Présentée par :

Basile JUMENTIER

Direction de thèse :

Olivier FRANCOIS — Directeur de thèse
professeur, Université Grenoble Alpes

Johanna LEPEULE — Co-directrice de thèse
Chercheur, Université Grenoble Alpes

Rapporteurs :

VIVIAN VIALLON
Maître de conférences HDR, UNIVERSITE LYON 1 - CLAUDE BERNARD

MATHIEU EMILY
Professeur, AGROCAMPUS- OUEST

Thèse soutenue publiquement le **9 juin 2022**, devant le jury composé de :

VIVIAN VIALLON Maître de conférences HDR, UNIVERSITE LYON 1 - CLAUDE BERNARD	Rapporteur
MATHIEU EMILY Professeur, AGROCAMPUS- OUEST	Rapporteur
VINCENT BONNETERRE Professeur des Univ. - Praticien hosp., UNIVERSITE GRENOBLE ALPES	Président
MAGALI RICHARD Chargé de recherche, CNRS DELEGATION ALPES	Examinatrice

Invités :

JOHANNA LEPEULE
Chargé de recherche HDR, INSERM DELEGATION AUVERGNE RHONE-ALPES

OLIVIER FRANCOIS

Table des matières

1. Introduction

Le tabagisme est une « intoxication aiguë ou chronique de nature physiologique et psychique provoquée par l'abus du tabac » (Encyclopédie médicale Quillet, 1965). Les fumées de cigarettes sont composées de plus de 4000 substances chimiques, dont une grande quantité sont reconnues toxiques. On sait également qu'au minimum 50 substances sont connues pour être cancérigènes (Institut National du Cancer, 2011). Le tabagisme a un effet néfaste sur la santé. Il est lié au développement de nombreux cancers (poumon, larynx, voies aérodigestives supérieures, bouche, vessie, œsophage, rein, pancréas, col utérin, sein et pénis). Il est également associé à de nombreuses maladies respiratoires, digestives et de la grossesse. Le tabagisme est la première source de mort évitable dans le monde (Barataud, 2016).

En France, on compte environ 11,4 millions de fumeurs, ce qui équivaut à 32% des 18-75 ans (Lermenier-Jeannet, 2014), dont 25% fument régulièrement. Le pourcentage de consommateurs réguliers de tabac diffère entre les hommes (35%) et les femmes (28%). Les hommes fumeurs consomment en moyenne 14 cigarettes par jour tandis que les femmes en consomment 12 par jour. A noter que la consommation de tabac diminue avec l'âge. Chaque année, on impute 78000 décès au tabagisme, soit environ 840 décès par semaine. Concernant les cancers, un cancer sur trois est lié au tabagisme et 90% des cancers du poumon ont pour origine le tabagisme. De plus, un fumeur régulier a une chance sur deux de mourir des conséquences de son tabagisme.

Depuis les années 70, la consommation de tabac a diminué au sein de la population française passant de 42% en 1974 à 32% en 2018. Malheureusement, cette diminution n'est effective que pour les hommes qui sont passés de 59% à 35% alors que les femmes ont quant à elles stagné à 28%.

Le tabagisme maternel prénatal peut provoquer des complications de la grossesse telles que le placenta preavia (Ananth et al., 1996), un risque de grossesse extra-utérine, un retard de croissance intra-utérine (Cnattingius, 2004), un risque de prématurité (Nabet et al., 2005), un risque du décès du nourrisson et des troubles respiratoires pour l'enfant à la naissance (Berlin, 2019; Peiffer et al., 2018).

En France, grâce aux Enquêtes Nationales Prénatales (ENP), on estime qu'en 2016, 17% des femmes enceintes fument au moins une cigarette par jour durant le troisième trimestre de grossesse (ENP2016) et 30% fument avant la grossesse. Ce pourcentage est identique à celui de 2010 (ENP2010).

L'analyse de l'ENP2016 (Blondel et al., 2017; Demiguel, 2018) a mis en évidence de nombreux effets néfastes du tabagisme sur la santé de la mère ainsi que sur celle de l'enfant. On peut noter que le risque de grossesse extra-utérine est augmenté pour les mères fumeuses et de manière dose dépendante avec le nombre de cigarettes journalières. Le risque de prématurité est augmenté chez les mères fumeuses mais ce risque disparaît si la femme met fin à son activité de fumeuse avant la conception.

On note également une augmentation dose dépendante du risque de fausse couche. Chez les fumeuses régulières (> 20 cigarette/jour), il existe un risque de 20% de fausse couche contre 10% pour des non fumeuses. Si la consommation de cigarettes est supérieure à 35 cigarettes par jour, le risque de fausse couche augmente jusqu'à 35%. Ces fausses couches sont principalement dues à des altérations induites par la nicotine présente dans le tabac.

Concernant le fœtus, l'un des premiers risques est l'augmentation du retard de croissance intra-utérin. On sait que le nombre de cigarettes consommées par jour va influencer négativement le poids à la naissance (Kataoka et al., 2018). Le tabagisme maternel serait lié à 11% des morts fœtales in utero (Leke et al., 2017). Il est également possible que le tabagisme puisse participer à des malformations du fœtus, mais aujourd'hui ce point est encore sujet à discussion.

La santé future des enfants de mères fumeuses peut également être altérée par le tabagisme maternel. Le risque de mort subite du nourrisson est multiplié par deux si la mère a fumé pendant la grossesse. Le tabagisme maternel favorise l'apparition de troubles respiratoires chez l'enfant tel que l'asthme et infections respiratoires (Burke et al., 2012; Peiffer et al., 2018). Le neurodéveloppement est altéré (Clifford et al., 2012). Une étude, portant sur 17000 enfants (Fogelman and Manor, 1988), montre que les enfants de mères fumant durant la grossesse obtiennent des scores de lecture, de mathématique et un niveau socio-professionnel plus faible que les enfants de mères non-fumeuses. De plus, les auteurs de cette étude ont mis en évidence l'apparition de troubles

du comportement chez les enfants de femmes fumant plus de 20 cigarettes par jour. Enfin, le tabagisme maternel prénatal peut engendrer l'apparition de maladies cardiaques (Cupul-Uicab et al., 2012; Ino, 2010) ou de cancers chez l'enfant (Chu, 2016; Rumrich et al., 2016).

Lors de la grossesse, les échanges entre la mère et le fœtus sont régis par le placenta. Une partie des constituants du tabac passent la barrière placentaire et va donc atteindre le fœtus.

Le placenta est un organe unique qui lie le fœtus à la mère de manière physique et biologique. Durant l'ensemble de la grossesse, il va apporter au fœtus les nutriments et l'oxygène nécessaires à son bon développement et évacuer les déchets. Il est aussi responsable de la production et sécrétion d'hormones qui assurent le bon déroulement de la grossesse pour la mère et le fœtus. Le placenta est donc un organe clé de la grossesse et du développement du fœtus. Il est important de noter que le placenta est porteur de l'ADN du fœtus. Des études montrent que les fonctions du placenta peuvent être altérées par le tabagisme (Appleton et al., 2013; Breton et al., 2014; Suter et al., 2011, 2010).

Avec l'avènement du séquençage haut débit il est devenu plus aisé d'avoir accès aux données omiques (ADN, ARN, méthylation, …) des individus. Nous nous intéresserons particulièrement à la méthylation de l'ADN car elle présente plusieurs intérêts : elle va agir sur l'expression des gènes, être impactée par les facteurs environnementaux tels

que le tabagisme (Rousseaux et al., 2020) ou la pollution atmosphérique (Abraham et al., 2018), et il s'agit à l'heure actuelle de la marque épigénétique la plus pertinente à analyser dans une étude épidémiologique. Les puces Illumina 27K, 450K ou 850K permettent de mesurer plusieurs centaines de milliers de marqueurs de méthylation. L'utilisation des statistiques va permettre de détecter et quantifier les relations entre les marqueurs d'intérêt, les variables d'expositions et les phénotypes. On parle alors d'études d'associations. Ces études d'associations seront décrites dans les parties 1.1.1 et 1.1.2. L'objectif de la thèse sera d'identifier des marqueurs de méthylations de l'ADN liés au tabagisme maternel prénatal mais aussi au poids de naissance de l'enfant, et qui pourraient ainsi expliquer la relation observée entre le tabagisme maternel et le poids de naissance de l'enfant. Le poids de naissance est connu pour être un marqueur robuste de la santé future. Un faible poids à la naissance est notamment associé à une augmentation du risque d'apparition de maladie cardiovasculaire et une fragilité sur la santé future (Rich-Edwards et al., 1997).

Pour identifier les différents marqueurs, l'outil statistique utilisé est l'analyse de médiation qui permet de quantifier la relation entre une exposition et un évènement de santé via l'inclusion d'une troisième variable médiatrice. Dans notre cas, l'exposition sera le tabagisme maternel prénatal, l'évènement de santé sera le poids de naissance et la variable médiatrice sera la méthylation de l'ADN placentaire. La grande dimension des données de méthylation de l'ADN engendre plusieurs problématiques statistiques notamment liées au fait que les analyses de médiation ont été développées pour une seule variable médiatrice et non un ensemble de médiateurs potentiels. Plusieurs travaux

récents ont proposé des analyses de médiation en grande dimension mais il n'existe pas de consensus sur la méthodologie la plus pertinente à utiliser (Blum et al., 2020; Zeng et al., 2021).

Quelques études se sont intéressées à la question de la médiation des effets du tabagisme maternel sur le poids de naissance via la méthylation de l'ADN mesurée dans le sang de cordon (Küpers et al., 2015; Xu et al., 2021).

Seules deux études (Cardenas et al., 2019; Morales et al., 2016) se sont focalisées sur la méthylation de l'ADN placentaire. L'équipe de Morales a travaillé sur la cohorte espagnole INMAA (n = 179) et a détecté deux médiateurs situés sur les gènes suivants : TRIO et entre les gènes *LINC00086* et *LEKR1*. L'équipe de Cardenas a travaillé quant à elle sur la cohorte canadienne GEN3G (n = 441) et a découvert 7 médiateurs situés sur les gènes suivants : *MDS2*, *PBX1*, *CYP1A2*, *VPRBP*, *WBP1L*, *CD28* et *CDK6*. D'autres études de médiation en haute dimension basées sur différentes méthodes ont été publiées dans différente domaines d'application. Néanmoins aucune comparaison de ces méthodes n'a été réalisée et aucun consensus n'existe quant à la méthode la plus pertinente à appliquer.

La thèse est articulée autour de deux grands chapitres qui sont précédés par un état de l'art. L'état de l'art porte à la fois sur les problématiques biologiques abordées, les méthodologies utilisées et la mise en évidence d'un besoin de nouvelles méthodes pour répondre à nos questions biologiques.

Le premier chapitre s'intéressera à la relation entre une exposition, ou un phénotype, et un ensemble de données de séquençage. Nous décrivons dans ce chapitre une nouvelle méthode, nommée Sparse LFMM, permettant de réaliser ces études d'associations. Nous avons ensuite comparé Sparse LFMM à un ensemble de méthodes de la littérature dans un contexte de simulation. Finalement, nous avons utilisé Sparse LFMM sur deux applications biologiques. La première porte sur la détection de gènes liés à un phénotype de floraison chez *Arabidopsis thaliana*. La seconde porte sur la compréhension de l'impact du tabagisme maternel prénatal sur la méthylation de l'ADN placentaire. L'enjeu de ce chapitre est donc de comprendre la relation entre une exposition (ou phénotype) et des données de séquençage.

Le deuxième chapitre s'intéresse quant à lui, à la compréhension de l'impact du tabagisme maternel prénatal sur la santé future de l'enfant via la méthylation de l'ADN. Pour se faire, nous avons développé une nouvelle méthode de médiation en haute dimension, nommée HDMAX2. Celle-ci a été testée puis comparée à un ensemble de méthodes équivalentes dans un contexte de simulations. Enfin, nous avons utilisé HDMAX2 pour estimer l'impact du tabagisme maternel prénatal médié par la méthylation de l'ADN placentaire sur le poids et l'âge gestationnel à la naissance.

1.1. Études d’associations

Nous nous intéressons au lien entre une matrice de données de séquençage et une exposition ou un évènement de santé. Il est donc important de comprendre les méthodologies existantes pour répondre à ces questions.

Les études d’associations sont des outils statistiques permettant de tester la relation entre un phénotype (ou une exposition) et un ensemble de marqueurs génétiques en grande dimension.

1.1.1. Études d’associations génomiques

Dans un premier temps les études d’associations ont été développées dans le cadre des « GWAS » pour Genome Wide Association Study (étude d'association pangénomique). Elles ont pour but de tester un ensemble de marqueurs génétiques contre un trait phénotypique. Elles permettent donc d’identifier des marqueurs génétiques significativement associés avec un phénotype dans une population donnée.

Dans la majorité des cas, ces études se focalisent sur les associations entre les polymorphismes nucléotidiques (SNP) et des phénotypes de maladies. Plus concrètement, ces études permettent d’identifier les variants génétiques les plus fréquemment associés avec un phénotype connu. A noter que les GWAS permettent de tester un très grand nombre de marqueurs génétiques à la fois (plusieurs millions).

D’un point de vue statistique, une GWAS consiste à réaliser un ensemble de régressions linéaires entre une matrice $M_{n,p}$ représentant les données de génétique et un vecteur $Y_{n,1}$ représentant un phénotype. Avec n le nombre d’individus, p le nombre de marqueurs génétiques et E une matrice d’erreur résiduelle :

$$Y = Ma^T + E$$

Dans une GWAS, on cherche donc à estimer les tailles d’effets $a_{1,p}$ et à tester leur significativité par des tests statistiques.

Toutefois, il existe un grand nombre de facteurs limitant les performances des GWAS. Premièrement la taille d’échantillon est souvent trop limitée pour réaliser des tests statistiques robustes. Une autre limitation est le nombre de tests utilisés qui est égale nombre de marqueurs génétiques utilisés. La population testée peut présenter une stratification, c’est-à-dire des fréquences alléliques différentes au sein d’une même population. De plus, il peut exister des facteurs de confusion connus ou non dans ces analyses. Enfin, il est intéressant de noter que les phénotypes analysés sont régulièrement polygéniques et portent sur des faibles tailles d’effets qui sont donc plus difficiles à identifier.

Pour pallier à la problématique de la structure des données génomiques, il faut ajuster les GWAS sur cette structure. Celle-ci n’étant pas connue, il est nécessaire de l’estimer. Une méthode consiste à déconvoluer la matrice de génotype ($M_{n,p}$) en deux matrices de rang K ($A_{n,K}$, $T_{K,p}$) :

$$M = AT$$

Ensuite, on peut ajuster la GWAS par la matrice *A*. Pour estimer cette matrice *A*, on peut extraire les composantes principales (Wang et al., 2019). Les méthodes telles que EIGENSRAT (Price et al., 2006), EIGENSOFT (Patterson et al., 2006) ou PLINK (Purcell et al., 2007) reprennent ce fonctionnement. Pour ce qui est des régressions, les plus utilisées pour réaliser les GWAS sont : les régressions linéaires, les régressions linéaires généralisées (GLM) et les modèles mixtes linéaires (LMM). Les GLM permettent de bien prendre en compte des facteurs de confusion connus tels que l'âge, le sexe, etc. Les LMM sont de puissants outils pour contrôler la covariance due à la structure de corrélation des données de génotype (Wang et al., 2019). Il existe bien d'autres méthodes ; on peut citer un modèle bayesien de LMM (BSLMM - Bayesian Sparse Linear Mixed Models) (Zhou et al., 2013; Zhou and Stephens, 2012), qui est une méthode hybride combinant des modèles linaires mixtes avec des régressions dites « sparse ».

Dans la majorité des cas, les tests d'associations des GWAS présentent une inflation de faibles *P*-valeurs qui va entraîner l'apparition de fausses découvertes. Pour pallier à ce problème, il est possible de calibrer les statistiques de tests par le « Genomic Inflation Factor » (GIF) (Devlin and Roeder, 1999).

1.1.2. Études d'associations épigénomiques

Contrairement au GWAS, les études d'associations épigénomiques (EWAS) s'intéressent au lien entre des marqueurs épigénomiques et une exposition ou un phénotype (Rakyan et al., 2011). De nombreuses études portent sur la méthylation de l'ADN.

La méthylation de l'ADN est un processus épigénétique qui correspond à l'addition de groupement méthyle sur certaines bases nucléotidiques. Chez l'homme, la méthylation s'effectue sur les cytosines présentes dans les séquences C-G (cytosine et guanine). On parle donc de dinucléotide CpG, pour cytosine–phosphate–guanine. Au sein du génome, les dinucléotides CpG ont une distribution particulière contrairement aux autres dinucléotides (GpC, ApT ou TpA), car ils forment des îlots de CpGs ayant une très forte concentration en CpGs. Ces îlots jouent un rôle dans la régulation de l'expression génétique (Jabbari and Bernardi, 2004).

De manière générale, la méthylation de l'ADN joue un rôle important dans de nombreux mécanismes cellulaires, notamment dans l'expression des gènes. Cependant son rôle est complexe, généralement une forte méthylation d'un CpG va entraîner une diminution de l'expression d'un gène. Mais cette action peut être différente selon la localisation du CpG sur le corps du gène, ou sur un promoteur du gène ou sur une région enhancer ou entre le site de transcription et le premier codon. A noter que la méthylation de l'ADN est considérée comme un mécanisme réversible mais les processus de déméthylation ne sont pas encore bien connus (Gkountela et al., 2015).

Un point très intéressant est que la méthylation de l'ADN est influencée par des facteurs environnementaux qu'ils soient sociaux, nutritionnels ou toxicologiques. De plus, la méthylation de l'ADN est fortement influencée par l'âge. La méthylation de l'ADN est donc un marqueur très intéressant car d'une part elle joue un rôle dans l'expression génique et d'autre part elle est impactée par les expositions de l'environnement.

La mesure de la méthylation de l'ADN est normalisée en « beta – value » et cette valeur varie entre 0 (hypométhylé) et 1 (hyperméthylé). Il existe une seconde valeur de méthylation la « M – value » qui correspond à une transformation logarithmique de la « beta – value ».

$$M = log_2(\frac{\beta}{1 - \beta})$$

Les études d'associations à l'échelle de l'épigénome sont récentes, la première date de 2011 ; c'est donc un nouveau domaine avec une méthodologie peu établie et en constante évolution. Les EWAS ont des limitations très similaires aux GWAS : taille d'échantillon souvent insuffisante, un grand nombre de tests statistiques et des facteurs de confusions parfois inconnus. L'une des limitations majeures est la prise en compte de la composition cellulaire des tissus testés. Cette composition, si elle n'est pas prise en compte ou mal estimée peut biaiser les résultats des EWAS en augmentant le nombre de faux positifs. La composition cellulaire est très rarement connue, il est donc nécessaire de l'estimer. Il existe un très grand nombre de méthodes permettant d'estimer cette composition cellulaire à partir de la matrice méthylation des « beta – value », matrice comportant n lignes (échantillons) et p colonnes (marqueurs CpGs).

Les deux méthodes les plus utilisées sont RefFreeEWAS (Houseman et al., 2016) (182 citations) et Refactor (Rahmani et al., 2016) (174 citations). Pour estimer la composition cellulaire, ces deux méthodes utilisent la déconvolution matricielle, RefFreeEWAS utilise une NMF (Non-negative Matrixe Factorization) tandis que Refactor utilise une décomposition en valeurs singulières (SVD) d'une analyse en composante principale (ACP). Ces deux méthodes ne considèrent pas l'ensemble des marqueurs pour optimiser leurs performances. Ainsi, pour RefFreeEWAS, il est conseillé de retirer les CpGs ayant une trop faible variabilité ou étant corrélé avec un facteur de confusion. De même, Refactor réalise la déconvolution uniquement sur les 500 CpGs les plus informatifs, pour se faire Refactor calcule un score d'importance relative pour chaque CpGs et ne conserve que les 500 CpGs ayant le plus grand score. A noter que ces méthodes ont besoin d'une information très importante qui est le nombre K de type cellulaire à estimer. Pour se faire, la méthode la plus simple et la plus robuste est la méthode dite du coude de l'ACP (méthode de Cattell), qui via la visualisation de l'éboulis des valeurs propres de ACP permet d'identifier le nombre de types cellulaires (Decamps et al., 2020).

En plus de RefFreeEWAS et Refactor, il existe un grand nombre de méthodes d'estimations de la composition cellulaire ; ces méthodes ressemblent fortement à RefFreeEWAS et Refactor dans leur procédure mais diffèrent par les méthodes de déconvolution utilisées ou les types de filtrations utilisés sur les CpGs. On peut citer les méthodes Edec (Onuchic et al., 2016) (citations 67) et Medecom (Lutsik et al., 2017) (61 citations). Il existe un grand nombre d'articles de synthèse (Decamps et al., 2020; Teschendorff and Zheng, 2017) discutant des performances de ces méthodes. Mais la

limitation majeure de ces méthodes est la non prise en compte des variables d'expositions dans l'estimation des compositions des types cellulaires (Caye et al., 2019).

C'est ici qu'intervient l'utilité des modèles à facteurs latents. Dans le cas des EWAS, les facteurs latents incluent donc les types cellulaires mais aussi possiblement des facteurs de confusions non observés. La prise en compte de ces facteurs de confusion non observés va permettre d'augmenter la performance des analyses statistiques.

La méthode pour réaliser des EWAS repose sur une régression entre la matrice de méthylation ($M_{n,p}$), l'exposition ($X_{n,1}$) et en incluant les compositions cellulaires et autres facteurs de confusions.

$$M = Xa^T + E$$

Ensuite, la significativité des tailles d'effets $a_{1,p}$ est évaluée via un test statistique tel que le test de Student par exemple. Le nombre de tests effectué correspond au nombre de marqueurs testés. Il faut donc utiliser une correction pour les tests multiples telle que la correction de Bonferroni ou le contrôle du taux de fausse découverte (FDR pour False Discovery Rate) ou le contrôle du taux d'erreur par famille (FWER pour Family-Wise Error Rate).

Un CpG est un marqueur portant sur une seule position chromosomique et dans l'optique de résonner par région chromosomique il est possible de détecter des DMRs (Differentially Methylated Regions). Un DMR est une région génomique ayant un statut de méthylation semblable tout au long de ces positions chromosomiques. Les DMRs

peuvent être impliqués dans la régulation transcriptionnelle des gènes. Les DMRs peuvent être détectés après l'analyse des CpGs par des méthodes telles que Comb-p. Cette méthode repose sur la correction de Stouffer-Liptak-Kechris qui combine les valeurs de significativité des CpGs adjacents dans des fenêtres glissantes.

1.2. Analyse de médiation

Les analyses de médiations ont été développées par les psychologues dans les années 80 pour comprendre la relation entre une variable indépendante et une variable dépendante via l'inclusion d'une troisième variable hypothétique appelée variable médiatrice ou simplement médiateur. En plus de la relation directe entre la variable indépendante et la variable dépendante, l'analyse de médiation propose un second chemin de causalité passant par la variable médiatrice. Ainsi, la variable dépendante va influencer la variable médiatrice qui va ensuite influencer la variable dépendante. Et donc l'analyse de la variable médiatrice va aider à mieux comprendre la relation entre la variable indépendante et dépendante (Baron and Kenny, 1986).

Lors d'une analyse de médiation, on cherche à estimer plusieurs effets (Figure 1.1) : l'effet total (*c*), l'effet direct (*c'*) et l'effet indirect (*ab*). L'effet total correspond à l'effet de la variable indépendante (*X*) sur la variable dépendante (*Y*). L'effet direct correspond à l'effet de la variable indépendante (*X*) sur la variable dépendante (*Y*) sachant la variable médiatrice. L'effet indirect correspond au produit des tailles d'effets *a* et *b*. L'effet a correspond à l'effet de la variable indépendante (*X*) sur la variable médiatrice (*M*) tandis que l'effet *b* correspond à l'effet de la variable médiatrice (*M*) sur la variable dépendante (*Y*) sachant la variable indépendante (*X*).

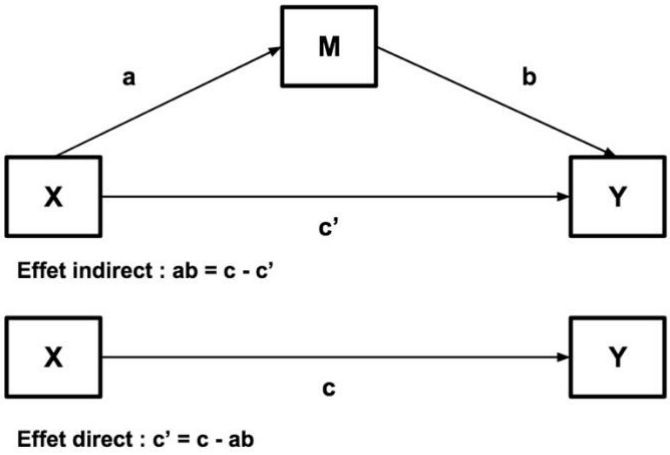

Figure 1.1 : Modèle de médiation*. Variable indépendante X, variable Médiatrice M et variable dépendante Y. Effet total c.*

L'enjeu méthodologique des analyses de médiation va être de tester la significativité de ces différents effets (*hypothèse H0 : effet = 0 ; hypothèse H1 : effet ≠ 0*). Pour ce qui est des effets *a*, *b*, *direct* (c') et *total (c)*, on peut facilement les estimer et les tester via l'utilisation de régression linéaire et de test de Student. Par contre, pour ce qui est de l'*effet indirect*, on peut facilement l'estimer mais il est beaucoup plus complexe de tester sa significativité. De nombreuses méthodes ont été développées pour répondre à cette problématique.

L'un des premiers chercheurs à avoir tenter de répondre à cette question est Michael Sobel. Il a mis au point un test statistique permettant de tester la significativité de l'effet indirect : le test de Sobel (Sobel, 1982). Pour se faire, Sobel estime l'erreur standard (SE) de l'effet indirect via la formule suivante :

$$SE = \sqrt{a^2\sigma_b^2 + b^2\,\sigma_a^2}$$

Où *a* représente l'effet de la variable indépendante sur la variable médiatrice et σ_a sa variance tandis que *b* représente l'effet de la variable médiatrice sur la variable dépendante sachant la variable dépendante et σ_b sa variance. Ensuite il peut déduire la statistique *t* de la façon suivante :

$$t = \frac{ab}{SE}$$

Où ab correspond à l'effet indirect. Pour finir, on détermine la significativité de l'effet indirect en comparant la statistique *t* à une distribution normale. Le test de Sobel est très facile à mettre en place et ne nécessite pas une forte capacité de calcul. Il a toutefois une faible puissance statistique ce qui le rend très conservatoire. Il est donc nécessaire d'avoir des échantillons de grande taille pour détecter des effets significatifs et ceci n'est pas toujours le cas dans les études observationnelles. Mackinnon (MacKinnon et al., 2002) a mis en place une règle empirique de bonne utilisation du test de Sobel. Ainsi, pour détecter un faible effet indirect, il est nécessaire d'avoir une taille d'échantillon égale à 1000 ; pour détecter un effet indirect moyen, une taille d'échantillon de 100 est nécessaire tandis que pour détecter un fort effet, 50 échantillons suffiront. Il existe des variantes du test de Sobel, tel que celui de Aroian (Aroian et al., 1978) ou de Goodman (Goodman, 1960) qui utilisent des statistiques *z* pour déterminer la signification ou qui calculent l'erreur SE légèrement différemment. Cependant, ces tests ont les mêmes limitations que le test de Sobel (MacKinnon et al., 2004).

Les seconds à s'être intéressés à cette question, sont les chercheurs Baron et Kenny qui ont développé une procédure en 3 étapes (Baron and Kenny, 1986). La première étape consiste à régresser la variable dépendante (*Y*) sur la variable indépendante (*X*) pour

confirmer que la variable indépendante est associée significativement à la variable dépendante. La seconde étape consiste à régresser la variable médiatrice (*M*) sur la variable indépendante (*X*) pour confirmer que la variable indépendante est significativement associée avec le médiateur. Enfin, la troisième étape consiste à régresser la variable dépendante à la fois sur le médiateur et sur la variable indépendante pour confirmer que le médiateur est associé significativement avec la variable dépendante. De plus, l'effet de la variable indépendante sur la variable dépendante doit être fortement réduit (voir nul) par rapport à l'effet identifié à l'étape 1. Cette méthode permet de détecter un médiateur mais il ne permet pas d'estimer l'effet indirect et donc encore moins de tester sa significativité. La procédure de Baron et Kenny a de plus été remise en cause par Hayes (Hayes, 2009) qui a démontré que la médiation peut exister en l'absence d'un effet total significatif et donc que l'étape 1 n'est pas nécessaire.

Au début des années 2000, un nouveau volet de méthodes se basant sur le bootstrap s'est développé pour tester l'effet indirect. La méthode de Preacher and Hayes (Preacher and Hayes, 2004) permet ainsi d'augmenter la puissance de détection des effets par rapport au test de Sobel. Ils utilisent le bootstrap, qui est une méthode de rééchantillonnage, pour estimer une approximation de la distribution de l'effet indirect. Cette méthode fournit donc les estimations et les intervalles de confiance permettant d'évaluer la significativité de la médiation. Les estimations révèlent la moyenne de l'effet indirect sur le nombre d'échantillons « bootstraper » et si zéro ne se situe pas dans les intervalles de confiance, on peut conclure que l'effet indirect est significatif. Étant donné qu'il faut de très nombreux rééchantillonnages (>1000) pour avoir des estimations, cette

méthode est coûteuse en temps de calcul, mais cela reste une des méthodes les plus utilisées à ce jour.

Les travaux d'Imai (Imai et al., 2010) vont dans le sens des méthodes utilisant le bootstrap. Lui et ses collaborateurs ont développé un package R « mediation » permettant de réaliser un grand nombre d'analyses de médiation unidimensionnelle. La particularité de ce package est qu'il permet de réaliser des analyses de médiation ayant des relations non linéaires. Ainsi, par exemple, il est possible d'utiliser des modèles linéaires généralisés ou des modèles additifs généralisés ainsi que d'autres modèles plus complexes. En plus des estimations de paramètres par bootstrap, il est possible d'estimer les paramètres par une méthode dite « quasi-Bayesian approximation ». De par sa grande polyvalence, ce package R est l'un des plus utilisés pour faire des analyses de médiation et il a notamment été utilisé par Cardenas (Cardenas et al., 2019) pour tester la médiation entre une exposition au tabagisme maternel, un sous ensemble de CpGs et le poids du bébé à la naissance.

Dans nos travaux, nous ne considérons pas un seul médiateur mais un très grand nombre de médiateurs potentiels (>300000). Et aucune des méthodes précédemment décrites n'a été développée pour prendre en compte un si grand nombre de médiateurs potentiels.

1.3. Analyse de médiation en haute dimension

La méthylation de l'ADN mesurée par les puces Illumina contient entre 27K et 850K CpGs. Ici, l'enjeu sera d'identifier les vrais médiateurs (Figure 1.2).

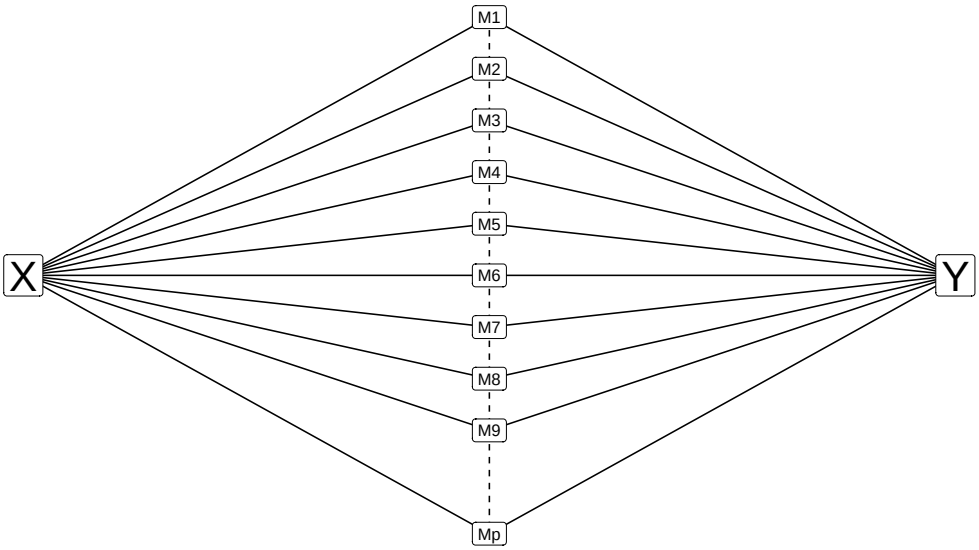

Figure 1.2 : Modèle de médiation en haute dimension*. Variable indépendante X, Médiateurs potentiels M_1..., Mp et variable dépendante Y.*

Il a été montré que les méthodes d'analyses de médiation classique, telles que les tests de Sobel ou de bootstrap, n'étaient pas adaptées à la grande dimension des données. Blum et al ont montré que le test de Sobel ne respectait pas l'hypothèse nulle en grande dimension car la distribution des *P*-valeurs n'est pas uniforme sous l'hypothèse nulle (Blum et al., 2020). De même, les méthodes utilisant le bootstrap ne sont pas adaptées à la grande dimension car elles montrent un manque de puissance et sont nécessitent un temps de calcul très long (Blum et al., 2020). Pour pallier ce problème, de nombreuses méthodes ont été développées permettant de considérer un grand ensemble de médiateurs potentiels. Ici, nous omettons volontairement les méthodes permettant de

prendre en compte un faible nombre de médiateurs (≈ 10). Il existe également un volet de méthodes développées sur des jeux de données possédant au maximum 1000 médiateurs potentiels. Ces méthodes permettent d'estimer et de tester l'ensemble des effets des médiateurs potentiels en prenant en compte la corrélation entre médiateurs (Zeng et al., 2021). Néanmoins, l'ensemble de ces méthodes ne permet pas de prendre en compte nos plus de 100000 médiateurs potentiels et nous ne les avons donc pas utilisés dans cette thèse. On peut tout de même citer les méthodes suivantes : Huang et Pan ont développé une méthode pour étudier les effets d'un micro ARN sur l'état de survie de patients atteints de glioblastome (Huang and Pan, 2016). Il existe aussi des méthodes bayésiennes : cmfilter (van Kesteren and Oberski, 2019) et bama (Song et al., 2020). L'application de cmfilter a permis d'identifier des CpGs faisant le lien entre les traumatismes de l'enfance et la réactivité au stress à l'âge adulte. A noter qu'ils n'ont pas pu considérer l'ensemble des CpGs séquencés (385 884) mais uniquement les 1000 CpGs les plus corrélés avec leur exposition et leur évènement de santé. Pour finir, comme pour cmfilter, bama a aussi été utilisé sur des données de méthylation d'ADN. Mais comme il est indiqué dans l'article correspondant, pour des raisons computationnelles les auteurs n'ont pu considérer que 2000 CpGs.

Aux vues des limitations computationnelles des méthodes ci-dessus, nous nous sommes focalisés sur les méthodes de médiation développées pour une très grande dimension (>100000). On remarque qu'elles ont été généralement développées pour des données de méthylations d'ADN (Dai et al., 2020; Djordjilović et al., 2020, 2019; Sampson et al., 2018; Zhang et al., 2016).

Les premiers, Zhang et al, ont développé HIMA (HIgh dimensional Mediation Analysis) (Zhang et al., 2016), composée de 3 étapes. La première étape consiste à réduire le nombre de médiateurs potentiels. Pour se faire, ils utilisent le « Sure Independence Screening » (SIS) (Fan and Lv, 2008) pour sélectionner les $\frac{n}{\log(n)}$ médiateurs potentiels les plus corrélés à la variable dépendante (*Y*) où *n* représente le nombre d'individu. La seconde étape consiste à sélectionner et estimer les effets des médiateurs en utilisant une régression pénalisée par la « pénalité concave minimax » (MCP) (Zhang, 2010). Pour finir, ils testent les effets estimés par la procédure MCP et appliquent la correction de Bonferroni pour les tests multiples. Il existe une variante d'HIMA développée par Gao (Gao et al., 2019) qui diffère uniquement au niveau de l'étape 2, dans laquelle les auteurs utilisent une pénalité lasso plutôt que MCP.

Les autres méthodes existantes sont différentes. A partir des valeurs de significativité (*P*-valeurs) de deux EWAS, elles vont tester l'hypothèse de médiation et contrôler pour les tests multiples. Les premières étapes de ces méthodes consistent à réaliser deux EWAS pour tester l'association entre la variable indépendante (*X*) et la matrice de médiateurs potentiels (*M*) puis l'association entre la matrice de médiateurs potentiels (*M*) et la variable dépendante (*Y*) sachant la variable indépendante (*X*).

$$M = Xa1^T + E$$

$$Y = Xa2^T + Mb1^T + E$$

De ces deux régressions, on s'intéresse uniquement aux valeurs de significativités (*Px* et *Py*) des tests des tailles des effets *a1* et *b1.* Ces méthodes vont donc prendre les deux séries de *P*-valeurs *Px* et P*y* et réaliser un test joint (*Joint Significance Test*). Le test joint

classique est de faire le maximum des deux séries de *P*-valeurs : $P = max\ (Px, Py)$. Ce test, comme le test de Sobel, évalue l'hypothèse nulle selon laquelle l'effet de l'exposition sur l'ADNm ou l'effet de l'ADNm sur l'évènement de santé est nul. Mais comme pour le test de Sobel, celui-ci n'est pas assez puissant en haute dimension et plusieurs méthodes l'ont amélioré.

ScreenMin (Djordjilović et al., 2020, 2019) propose une procédure en deux étapes pour contrôler le taux d'erreur par famille (FWER) du test joint. Pour se faire, leur première étape est de considérer une *P*-valeur minimale et ensuite ils appliquent le test joint classique :

- Étape 1 : Considérant un ensemble de *P*-valeurs minimales : {min P_1, ..., min P_m}, où min P_i = min {Px_i, Py_i}. Pour un c ∈ (0,1) donné, soit S = {i : min P_i ≤ c} un ensemble d'indices d'hypothèses sélectionnées.
- Étape 2 : Considérant un sous-ensemble de *P*-valeurs maximales {max P_i : i ∈ S}, et corrigez-les pour les tests multiples par la correction de Bonferroni ou Holm. *P*-valeurs ajustées notées max P^*_i. La *P*-valeur ajustée est définie par :

$$P_i^* = \begin{cases} \max P_i^* & i \in S \\ 1 & i \notin S \end{cases}$$

On peut citer une seconde méthode, HDMT (Dai et al., 2020), qui permet de produire des *P*-valeurs calibrées à partir du test joint classique (*Pmax*). Cette méthode construit une distribution nulle en utilisant une transformation quadratique des *P*-valeurs *Pmax* afin de corriger le FDR et le FWER.

Enfin, Sampson (Sampson et al., 2018) propose une procédure de comparaison multiple qui permet de rejeter un ensemble d'hypothèses nulles et donc d'identifier les médiateurs. En outre, comme les précédentes méthodes, ils contrôlent le FWER ou le FDR.

Toutes ces méthodes ont été publiées dans des journaux de méthodes, mais d'autres méthodes ont été directement appliquées à des analyses de données et publiées dans des journaux de différents domaines.

Ainsi, Morales (Morales et al., 2016), pour répondre à une problématique similaire à la nôtre, ont tout d'abord réalisé une EWAS, entre la variable indépendante (tabagisme maternel) et les médiateurs potentiels (méthylation de l'ADN), pour sélectionner un sous ensemble de médiateurs potentiels. Ils ont ensuite utilisé le test de Sobel pour tester la médiation du sous ensemble de médiateurs.

Sur la même question biologique, une autre équipe, celle de Cardenas (Cardenas et al., 2019), a utilisé une approche différente. Ils ont sélectionné le sous-ensemble de médiateurs potentiels en réalisant les deux EWAS, la première entre tabac et méthylation et la seconde entre méthylation et poids de naissance. Puis, ils ont utilisé le package d'Imai « mediation » pour tester la médiation.

Enfin, Tobi et al (Tobi et al., 2018), ont utilisé une approche originale pour sélectionner un sous ensemble de médiateurs potentiels parmi une matrice de données de méthylation dans leur étude de l'association entre l'adversité prénatale et les facteurs de

risque de maladie métabolique à l'âge adulte. Pour se faire, pour chaque médiateur potentiel, ils ont réalisé deux GEE (Generalized Estimating Equation), l'un où le médiateur potentiel est expliqué uniquement par la variable indépendante et l'autre où le médiateur potentiel est expliqué par la variable indépendante et dépendante. Ensuite, ils comparent les deux GEE à l'aide d'une ANOVA et utilisent la *P*-valeur de l'ANOVA comme *P*-valeur de médiation. A noter qu'ensuite ils estiment les effets indirects en utilisant le package « mediation ».

La majorité de ces méthodes sont très récentes excepté HIMA qui date de 2016. A ce jour, il existe peu d'articles comparant les méthodes énoncées (Zeng et al., 2021) mais aucun article n'a testé ces méthodes sur des simulations pour connaître réellement leurs performances. Il n'y a pour l'instant aucun consensus sur la méthode à utiliser lors de l'analyse de données réelles. De plus, comme nous l'avons vu, ces méthodes ne sont pas forcément utilisées par les chercheurs pour répondre à des questions biologiques. Aucune de ces méthodes ne s'intéressent réellement à un des problèmes fondamentaux des études d'associations, qui est la possible présence de facteurs de confusion non observés. Aux vues de ces limitations, il paraît raisonnable de développer une nouvelle méthode de médiation en haute dimension permettant, d'une part de prendre en compte des facteurs de confusions non observés et de tester la médiation en tenant compte de la problématique des tests multiples, et d'autre part de comparer ces méthodes dans un contexte de simulation pour estimer leurs performances.

1.4. Cohorte mère-enfant EDEN

Notre question biologique se concentre sur la compréhension de l'impact du tabagisme prénatal sur la santé future de l'enfant via la méthylation de l'ADN. Nous avons montré que cette question biologique nécessite le développement de nouvelles méthodes statistiques. Pour répondre à notre problématique biologique nous avons utilisé un jeu de données issu de la cohorte EDEN (Heude et al., 2016).

La cohorte mère-enfant EDEN (pour étude des déterminants pré et postnataux du développement de la santé de l'enfant) a débuté en 2003 et a pour l'objectif de comprendre les facteurs de début de vie pouvant influer le développement de l'enfant à court, moyen et long terme. De nombreux facteurs sont examinés au sein de la cohorte, les plus importants sont la nutrition de la mère et de l'enfant, des marqueurs génétiques et épigénétiques, des expositions environnementales (tabac, polluants de l'air) et des facteurs sociaux. Pour cette étude, 2002 femmes enceintes, suivies dans les maternités des CHU de Nancy et Poitiers, ont été incluses dans la cohorte entre 2003 et 2005. Les données des familles ont été collectées par questionnaires, par examen médical ou via les dossiers de santé (Tableau 1.1). Les familles ont été suivies de manière régulière, et le sont toujours à ce jour. Suite à l'accouchement, de nombreux échantillons biologiques ont été recueillis (Tableau 1.2) tel que le placenta ou le sang de cordon.

Tableau 1.1 : Collecte des données de la cohorte mère-enfant EDEN. *Depuis (Heude et al., 2016).*

	24–26 WA	Delivery	4 months	8 months	1 year	2 years	3 years	4 years	5–6 years
Health									
Weight	♟ ✍	♟	–	–	♟	–	♟	–	♟
Height	♟	–	–	–	–	–	–	–	♟
Body composition (BIA)	–	–	–	–	–	–	–	–	♟
Blood pressure	♟	♟	–	–	–	–	–	–	♟
Heart rate	♟	♟	–	–	–	–	–	–	♟
Asthma/allergy	✍	✍	–	–	–	–	–	–	–
Infectious disease	✍	❐	–	–	–	–	–	–	–
Mental health	✍	♟	✍	✍	✍	✍	✍	✍	✍
Exposures									
Tobacco smoking	✍	✍	✍	✍	✍	✍	✍	✍	✍
Passive smoking	✍	✍	✍	✍	✍	✍	✍	✍	✍
Alcohol consumption	✍	✍	✍	✍	✍	✍	✍	–	–
Binge drinking	✍	✍	–	–	–	–	–	–	–
Substance misuse	✍	✍	–	–	–	✍	–	–	–
Medicine intake	–	✍	✍	✍	✍	✍	✍	–	–
Diet and dietary behaviour	✍	✍	–	–	–	✍	–	–	–
Stress/depression	✍	–	–	–	–	–	✍	–	✍
Occupational hazards	✍	✍	–	–	–	–	–	–	–
Outdoor air pollution	✍	✍	–	–	✍	✍	✍	–	–
Indoor contaminants	–	✍	–	–	–	–	–	–	–

The table specifies whether information was not collected at this follow-up (–), collected through health records (❐), self-administered questionnaires (✍) or from the midwives during clinical examinations (♟).

BIA, bio impendance analysis; WA, weeks of amenorrhoea. At 24–26 WA, many questions also concerned the pre-pregnancy period: on weight before pregnancy, weight evolution since age 20 years, diet during the year preceding pregnancy, alcohol consumption, tobacco use before and at the beginning of pregnancy and health antecedents.

Tableau 1.2 : Échantillons biologiques de la cohorte mère-enfant EDEN. *Depuis (Heude et al., 2016).*

	During pregnancy	*Birth*	*5–6 years*
DNA	Mother	Father, Cord	Child
Plasma	Mother	Father, Cord	Child
Serum	Mother (Fasting and 1 h post charge)	Cord	Child
Platelets	–	Cord	–
Urine	Mother	–	Child
Erythrocytes	Mother	Cord	–
Colostrum	–	Mother	–
Cord samples	–	Cord	–
Meconium	–	Child	–
Placenta samples	–	Placenta	–
Hair[a]	Mother	Mother, Child	Child
Saliva	Mother	–	–

The table specifies whether information was not collected at this follow-up (–).

[a]also collected at 3 years.

La méthylation de l'ADN placentaire a été mesurée grâce à des puces Illumina 450 K. Il a été mis en évidence que le tabagisme maternel prénatal altérait la méthylation de l'ADN et notamment dans les régions enhancers (Rousseaux et al., 2019). Une seconde étude

(Abraham et al., 2018), qui avait pour but de comprendre l'impact des polluants atmosphériques durant la grossesse, sur la méthylation placentaire a montré que des marqueurs situés sur le gène *ABORA2B* sont associés à la pollution au dioxyde d'azote (NO2). Il est intéressant de noter que ce gène est associé à la pré-éclampsie, une pathologie de la grossesse connue pour être également associée à l'exposition aux polluants de l'air.

1.5. Objectif de la thèse

La thèse s'articule autour de deux chapitres, répondant aux objectifs suivants.

Le premier chapitre vise à comprendre la relation entre une variable d'intérêt, tel une exposition ou un phénotype, et un ensemble de variables issues du séquençage, tel que des génotypes ou de la méthylation de l'ADN. On veut notamment résoudre les problèmes suivants : prise en compte de la structure des données omiques ainsi que la présence de facteur de confusion non connu ; et faire de la sélection de variable sans à avoir à utiliser de tests statistiques.

Au sein du second chapitre, on veut quantifier les relations entre le tabagisme maternel prénatal, la méthylation de l'ADN placentaire et la santé future de l'enfant. Pour ce faire, nous voulions développer une nouvelle méthode médiation en haute dimension permettant de répondre aux problématiques suivantes : prise en compte de la structure des données de méthylations ainsi que la présence de facteur de confusion non connu ; identification de marques de méthylations médiant la relation entre le tabagisme maternel prénatal et la santé future de l'enfant ; et quantification de l'effet du tabagisme maternel prénatal sur la santé future de l'enfant passant par la méthylation de l'ADN.

1.6. Résultats principaux

Avant d'aborder les chapitres de la thèse en détails, cette section résume les principales contributions de la thèse.

La thèse s'appuie sur deux publications

- Caye, K., Jumentier, B., Lepeule, J., & François, O. (2019). LFMM 2 : fast and accurate inference of gene-environment associations in genome-wide studies. *Molecular Biology and Evolution*, 36(4), 852-860. *Disponible en annexe.*

- Jumentier, B., Caye, K., Heude, B., Lepeule, J., & François, O. (2022). Sparse latent factor regression models for genome-wide and epigenome-wide association studies. Accepté par *Statistical Applications in Genetics and Molecular Biology*.

Et un article en cours de finalisation correspondant à la contribution principale de la thèse

- Jumentier, B., Barrot, C-C., Estavoyer, M. Heude, B., François, O. & Lepeule, J. HDMAX2 : HDMAX2 : A framework for High Dimensional Mediation Analysis with application to maternal smoking, DNA methylation and birth outcomes.

Dans un premier temps nous parlerons des méthodes présentes dans le package R LFMM développé par notre équipe et publié sur CRAN (https://cran.r-project.org/web/packages/lfmm/index.html). Et dans un second temps, nous aborderons le sujet de la médiation en haute dimension, via le développement d'une nouvelle méthode ainsi que son application sur les données de la cohorte mère-enfant EDEN. Cette application est l'objectif principal de la thèse qui est de caractériser l'impact du

tabagisme maternel sur le poids à la naissance et d'identifier des marques de méthylation liées au tabagisme et au poids de naissance. Dans cette partie, seuls les principaux résultats sont discutés et les études complètes seront plus amplement développées dans le chapitre 2 pour la valorisation des méthodes du package LFMM et dans le chapitre 3 pour l'étude de la médiation en haute dimension.

La première partie de la thèse s'est focalisée sur les études d'associations, c'est-à-dire étudier l'impact d'une exposition sur un ensemble de marqueurs génétiques ou alors étudier l'impact de marqueurs génétiques sur un phénotype. Plus précisément nous nous sommes appliqués à tester les méthodes présentes dans le package R LFMM. Ce package contient deux méthodes, ridge LFMM et sparse LFMM. Ces deux méthodes permettent de réaliser des études d'associations sur n'importe quel type de données, tel que des polymorphismes génétiques ou des marques de méthylation d'ADN. Ces méthodes ont été développées pour répondre à un problème bien connu dans les études d'associations : la présence de facteurs de confusion non observés dans les données. Les facteurs de confusion sont des variables corrélées à la fois avec la variable d'intérêt (exposition ou phénotype) et avec les marqueurs génétiques. Les facteurs de confusion peuvent être responsables de l'apparition de faux positifs dans les études d'associations et il est donc important de les prendre en compte. Ces facteurs peuvent être connus, par exemple l'âge (ou le sexe) qui est un facteur de confusion connu pour les données de méthylation d'ADN. Mais ils peuvent aussi être inconnus comme par exemple la composition des types cellulaires pour les données de méthylation d'ADN ou le taux de parenté pour des données de polymorphismes nucléotidiques. Dans ce contexte, ridge

LFMM et sparse LFMM proposent d'estimer des facteurs latents pour pallier à ce problème. Ces deux méthodes sont basées sur des modèles mixtes à facteurs latents et utilisent des algorithmes des moindres carrés régularisés pour estimer les facteurs latents et les tailles d'effets des marqueurs. Ridge LFMM utilise une pénalité de régularisation L^2 tandis que sparse LFMM utilise une pénalité L^1 et ces facteurs latents sont contraints par une norme nucléaire. La pénalité L^1 de sparse LFMM, permet de fixer une grande partie des tailles d'effets à 0 et ainsi de faire une sélection de variable sans avoir à effectuer de tests de statistiques qui sont nécessaires à ridge LFMM pour faire de la sélection de variable. Des études de simulations ont permis de démontrer les bonnes performances de ces deux méthodes par rapport à des méthodes présentes dans la littérature.

Les performances de sparse LFMM ont été évaluées grâce à deux études de simulation, la première utilisant un simulateur de données basé sur un modèle génératif tandis que la seconde utilise des simulations réalisées à partir de données réelles. Sparse LFMM a été comparée à un ensemble de méthodes. Sparse LFMM a été comparé à un ensemble de méthodes similaires : des méthodes dites "sparse", LASSO (Friedman et al., 2010) et BSLMM (Zhou et al., 2013; Zhou and Stephens, 2012), qui comme sparse LFMM permettent la sélection de variable via l'estimation d'un grand nombre de tailles d'effets nuls, et des méthodes dites "non-sparse", telles que ridge LFMM (Caye et al., 2019), CATE (Wang et al., 2017) et SVA (Leek et al., 2012).

Pour la première campagne de simulation, nous avons développé un simulateur mimant des données de méthylation d'ADN et d'exposition. De plus, pour évaluer les méthodes

dans un contexte de facteurs de confusion non-observés, nous avons inclus six facteurs de confusion dans nos simulations. Dans les simulations, nous faisons varier deux paramètres importants, la taille des effets des marques de méthylation causales et l'intensité de la confusion (part de variance expliquée par les facteurs non-observés). Pour chaque méthode, nous estimons les tailles d'effet des marques de méthylation. Pour estimer leurs performances, nous calculons la précision, le rappel et le F1-score pour les tops hits. Cette campagne de simulation a permis de mettre en évidence les bonnes performances de sparse LFMM par rapport aux autres méthodes « sparse ».

Pour notre seconde campagne de simulations, notre simulateur utilise des données réelles pour simuler un phénotype (François and Caye, 2018). Nous avons utilisé les données de polymorphismes nucléotidiques (SNP) du chromosome 5 d'*Arabidopsis thaliana*. Les simulations de phénotypes incorporent des caractéristiques réalistes telles que des interactions gène-environnement (GxE) à l'aide de variables environnementales extraites d'une base de données bioclimatiques. Dans ces simulations, nous avons fait varier les tailles d'effet des SNP causaux ainsi que l'intensité de l'interaction GxE qui est un facteur de confusion. Pour chaque méthode, nous estimons les tailles d'effet des marques de méthylation. Pour estimer les performances, nous calculons la précision, le rappel et le F1-score sur les tops hits. Ces simulations ont permis de mettre en évidence la robustesse de sparse LFMM par rapport aux autres méthodes testées.

Comme pour sparse LFMM, nous avons évalué les performances de ridge LFMM grâce à deux études de simulation, l'une utilisant un modèle de simulation générative et l'autre

utilisant des données réelles pour simuler une exposition. Ridge LFMM a été comparé à un ensemble de méthodes utilisées pour les études d'associations (PCA, CATE, SVA et t.test pour un test naïf). Nous intéressons à la performance des tests statistiques pour identifier des marqueurs associés à la variable d'intérêt. Nous effectuons pour cela des régressions linéaires entre l'exposition et la matrice de marqueurs en incluant en co-variables les facteurs latents estimés par chaque méthode. Puis nous testons les estimations des tailles d'effet de chaque marqueur pour obtenir une valeur de significativité (*P*-valeur) pour chaque marqueur. Enfin, pour estimer les performances de chaque méthode, nous calculons la précision, le rappel et la F1-score sur une liste de marqueurs respectant un seuil prédéfini de significativité.

Dans la première campagne de simulation, nous avons développé un simulateur de données simulant uniquement des données artificielles grâce à un modèle génératif. Ce simulateur nous permet de simuler une exposition, une matrice de marqueurs ainsi qu'un ensemble de facteurs de latents corrélés à la fois avec l'exposition et un sous ensemble de marqueurs. Les facteurs latents simulés ont pour objectif de miner des facteurs de confusion non observés. Nous avons testé quatre scénarios de simulations en faisant varier uniquement l'intensité de confusion qui correspond à la corrélation entre les facteurs latents, l'exposition et les marqueurs causaux.

La seconde campagne de simulation utilise des données réelles pour simuler une exposition. Pour respecter la chaîne de causalité, l'exposition est simulée de façon conditionnelle à un sous ensemble de marqueurs. Nous avons utilisé les données de

méthylation d'ADN placentaire de la cohorte EDEN (Heude et al., 2016) pour simuler différentes expositions. Quelle que soit la campagne de simulations, ridge LFMM a obtenu les meilleurs résultats F1-score ou précision.

Pour finir de tester sparse LFMM, nous l'avons appliquée à deux jeux de données réelles. En premier lieu nous avons réalisé une GWAS pour chercher des gènes liés à la floraison chez *Arabidopsis thaliana*. Cette étude s'appuie sur les données publiées de Atwell (Atwell et al., 2010). Nous avons pu mettre en évidence des gènes déjà connus pour être liés à la floraison (FLC et DOG1). Sparse LFMM a aussi permis de faire de nouvelles découvertes (ACL5, SAP).

Ensuite, nous avons appliqué sparse LFMM sur la cohorte EDEN, dans l'optique de comprendre l'impact du tabagisme sur la méthylation de l'ADN placentaire. Cette EWAS a permis de mettre en évidence plusieurs résultats biologiques. En plus de détecter un grand nombre de marqueurs liés au tabac, sparse LFMM a permis de montrer un enrichissement en régions amplificatrices (enhancer) et un appauvrissement en régions promotrices. De tels résultats avaient déjà été découverts sur la cohorte EDEN (Rousseaux et al., 2019).

Dans la GWAS et dans l'EWAS les résultats obtenus par sparse LFMM étaient concordants avec les résultats obtenus par ridge LFMM, CATE ou SVA.

Dans le second chapitre, nous nous intéressons à l'objectif principal de la thèse : la compréhension de l'impact du tabagisme maternel sur le poids à la naissance via la méthylation de l'ADN placentaire. L'état de l'art à mis en évidence le manque de méthodologie et qu'il n'existe pas de consensus pour réaliser des études de médiation en haute dimension. Ces études de médiation en haute dimension sont nécessaires à l'identification de marqueurs potentiellement médiateurs de la relation causale entre le tabagisme maternel et le poids à la naissance. C'est dans ce contexte que nous nous sommes penchés sur le développement d'une méthode de médiation permettant de sélectionner un sous-ensemble de médiateurs potentiels au sein de données de méthylation d'ADN. Notre méthode présente plusieurs avantages, d'une part elle permet la prise en compte de facteurs de confusion inconnus, de filtrer l'ensemble des marques de méthylation (CpGs), de sélectionner uniquement les potentiels médiateurs et pour finir c'est la seule méthode, à ce jour, à réaliser de la médiation sur des régions de CpGs, que nous appelons *régions médiatrices agrégées* (AMRs).

Nous avons ainsi développé une nouvelle procédure de médiation en haute appelée HDMAX2 (High Dimensional Mediation Analysis). Cette procédure comporte 2 étapes essentielles et combine une méthode d'EWAS (LFMM) et un test de médiation (max^2).
L'étape 1 consiste à réaliser deux EWAS, la première entre l'exposition et la matrice de méthylation et la seconde entre la matrice d'exposition et l'évènement de santé. L'étape 2 consiste à tester l'hypothèse de médiation via un test joint réalisé sur les *P*-valeurs des deux EWAS. Ici, nous introduisons un nouveau test joint : le test du maximum au carré (max^2).

Notre procédure permet d'identifier des régions médiatrices grâce à l'utilisation de la méthode de détection de DMRs comb-p (Xu et al., 2016). Pour finir, au sein de notre procédure nous introduisons une nouvelle quantité à estimer : l'effet total indirect (Overall Indirect Effect, OIE) qui représente l'effet global médié par l'ensemble des médiateurs confirmés.

Nous avons validé les deux étapes de notre procédure via des simulations et nous les avons comparées aux méthodes présentes dans la littérature. Tout d'abord nous avons développé un simulateur de données de médiation en grande dimension. Le simulateur de données est basé sur un modèle à facteur latent. Celui-ci sera décrit plus précisément dans la partie 3.2.

Nous avons comparé l'étape 1 de HDMAX2 à deux méthodes : RefFreeEWAS (Houseman et al., 2016) et Refactor (Rahmani et al., 2016). Les résultats ont permis de mettre en évidence la supériorité de HDMAX2 par rapport à ces méthodes. Pour la seconde étape de HDMAX2, les simulations ont montré que notre test joint max^2 ainsi que le test HDMT (Dai et al., 2020) obtiennent les meilleures performances mais nous avons montré que notre procédure a un temps d'exécution beaucoup plus court que celui d'HDMT. Grâce à ces campagnes de simulation, nous avons pu montrer que notre méthode montre de bonnes performances statistiques tout en étant la plus rapide.

Après le développement et la validation de notre procédure de médiation en haute dimension, nous avons pu nous intéresser à notre principale problématique biologique. Nous avons donc utilisé HDMAX2 sur la cohorte mère-enfant EDEN (Heude et al., 2016) pour comprendre l'impact du tabagisme maternel prénatal sur la santé future de l'enfant via la méthylation. Notre analyse porte sur 470 femmes et 379 904 CpGs. L'utilisation de HDMAX2 nous a permis de faire de nombreuses découvertes. Premièrement, HDMAX2 nous a permis de détecter de nombreux marqueurs (32 CpGs) médiant cette relation mais surtout de détecter des régions (19 AMRs) liées à la fois au tabagisme et aux résultats de santé de l'enfant. A noter que de nombreux marqueurs sont situés sur des gènes (*NECTIN1*, *AHR*, *FGFR2*, *COASY*, *BLCAP*, *SKI*, *AJAP1* et *SH3BP5)* liés à la pré-éclampsie, une maladie rare mais grave qui survient au cours de la grossesse et provoque de l'hypertension artérielle. Cette pathologie est liée à un tiers des naissances très prématurées et donc la surreprésentation des gènes liés à la pré-éclampsie soutient un effet indirect de ces marqueurs dans cette pathologie, ainsi qu'un effet du tabagisme.

Au sein des marqueurs détectés, nous avons pu mettre en évidence un enrichissement en régions enhancers et un appauvrissement en régions promotrices (Rousseaux et al., 2020).

Nous avons identifié une relation complexe entre le tabagisme maternel prénatal, la méthylation de l'ADN placentaire, l'âge gestationnel et le poids de naissance. Cette découverte impliquerait un phénomène de causalité inverse entre l'âge gestationnel, les AMRs situés sur les gènes *COASY* et *BLCAP* et le poids à la naissance. A noter qu'une

précédente étude de la cohorte EDEN (Abraham et al., 2018), qui portait sur l'impact des polluants atmosphériques sur la méthylation placentaire avait déjà mis en évidence la pré-éclampsie via le gène *ADORA2B.* De plus, le gène *BLCAP* a déjà été identifié comme étant lié au tabagisme maternel dans une EWAS effectuée sur la cohorte EDEN (Rousseaux et al., 2020).

Enfin, un des résultats intéressant de cette thèse est la part d'effet médié par la méthylation de l'ADN dans nos analyses de médiation. Ainsi, nous avons estimé que dans l'analyse entre le tabac et le poids à la naissance, l'effet passant par la méthylation de l'ADN est de l'ordre de -44g qui représente une part importante de l'effet total (-140g).

Cependant, nous n'avons pas été en mesure de détecter les gènes identifiés par les précédentes analyses effectuées par Morales et Cardenas (Cardenas et al., 2019; Morales et al., 2016). Concernant Morales, ce résultat peut s'expliquer par le fait que dans leur EWAS (tabac - méthylation), ils n'estiment pas de facteurs de confusions ou de compositions cellulaires. De plus, comme test de médiation, ils utilisent le test de Sobel, qui n'est pas adapté à la grande dimension des données de méthylations d'ADN (Blum et al., 2020). Pour ce qui est de Cardenas, ils utilisent Refactor dans leurs EWAS, pour estimer la composition, et via une étude de simulation nous avons montré que cette méthode obtient de faibles performances.

Tous les résultats présentés ci-dessus seront plus précisément décrits dans les chapitres suivants.

2. Sparse latent factor regression models for genome-wide and epigenome-wide association studies

Ce chapitre reprend la trame de l'article « Sparse latent factor regression models for genome-wide and epigenome-wide association studies » écrit par Jumentier Basile, Caye Kévin, Heude Barbara, Lepeule Johanna et François Olivier. L'article a été accepté par la revue « Statistical Applications in Genetics and Molecular Biology ».

Résumé :

L'association de phénotypes ou d'expositions avec des données génomiques et épigénomiques se heurte à d'importants défis statistiques. L'un de ces défis consiste à tenir compte de la variation due à des facteurs de confusion non observés, tels que l'ascendance individuelle ou la composition du type cellulaire dans les tissus. Ce problème peut être résolu avec des modèles de régression à facteurs latents pénalisés, où des pénalités sont introduites pour faire face à une dimension élevée dans les données. Si une proportion relativement faible de marqueurs génomiques ou épigénomiques est corrélée à la variable d'intérêt, les pénalités de parcimonie (sparse) peuvent aider à capturer les associations pertinentes, mais l'amélioration par rapport aux approches non parcimonieuses n'a pas encore été entièrement évaluée. Ici, nous présentons des algorithmes des moindres carrés qui estiment conjointement les tailles d'effet et les facteurs de confusion dans des modèles de régression à facteurs latents parcimonieux (sparse). Dans les données simulées, les modèles de régression à facteurs latents parcimonieux ont généralement obtenu des performances statistiques plus

élevées que les autres méthodes parcimonieuses, tel que LASSO (Least Absolute Shrinkage and Selection Operator) et BSLMM (Bayesian Sparse Linear Mixed Model). Dans les simulations de modèles génératifs, les performances statistiques étaient légèrement inférieures (tout en étant comparables) aux méthodes non parcimonieuses, mais dans les simulations basées sur des données empiriques, les modèles de régression à facteurs latents parcimonieux étaient plus robustes que les approches non parcimonieuses. Nous avons appliqué des modèles de régression de facteurs latents parcimonieux à une étude d'association à l'échelle du génome d'un trait de floraison de la plante *Arabidopsis thaliana* et à une étude d'association à l'échelle de l'épigénome du statut tabagique chez les femmes enceintes. Pour les deux applications, les modèles de régression à facteurs latents parcimonieux ont facilité l'estimation des tailles d'effet non nulles tout en surmontant les problèmes de tests multiples. Les résultats étaient non seulement cohérents avec les découvertes précédentes, mais ils ont également identifié de nouveaux gènes avec des annotations fonctionnelles pertinentes pour chaque application.

2.1. Introduction

Les études d'association sont l'un des outils les plus puissants pour identifier les variations génomiques liées à des pathologies, à des expositions ou des phénotypes. Ces études sont divisées en plusieurs catégories en fonction de la nature des données génomiques des marqueurs. Par exemple, les études d'association à l'échelle du génome (Genome Wide Association Study ou GWAS) se concentrent sur le polymorphisme d'un seul nucléotide chez différents individus pour estimer les effets des allèles sur la maladie (Balding, 2006), tandis que les études d'association à l'échelle de l'épigénome (Epigenome Wide Association Study ou EWAS) mesurent l'association entre des marques épigénétiques, telles que les niveaux de méthylation de l'ADN, et une exposition ou un phénotype (Rakyan et al., 2011).

En dépit de leur succès pour identifier des marqueurs liés à une exposition ou phénotype, les études d'association peuvent être faussées par le problème de confusion, qui se pose lorsque des variables non-observées sont simultanément corrélées avec les variables d'intérêt (exposition ou phénotype) et les marqueurs génomiques (Wang et al., 2017). Les approches historiques du problème de confusion tiennent compte des facteurs de confusion non observés en considérant des corrections pour l'inflation (Devlin and Roeder, 1999) et en utilisant des méthodes empiriques de test de l'hypothèse nulle (Efron, 2004).

Des approches alternatives évaluent les facteurs de confusion non observés en utilisant des combinaisons linéaires de variables observées. Dans les GWAS, une approche fréquemment utilisée, consiste à calculer les composantes principales de la matrice de génotype, et à les inclure comme covariables dans des modèles de régression linéaire (Price et al., 2006). La variable d'intérêt peut cependant être colinéaire aux premières composantes principales, et la suppression de leurs effets peut entraîner une perte de puissance statistique. Pour surmonter ce problème et augmenter la puissance statistique, des méthodes basées sur des modèles de régression à facteurs latents ont été proposées (Carvalho et al., 2008; Leek and Storey, 2007). Les modèles de régression à facteurs latents utilisent des méthodes de déconvolution dans lesquelles des variables non observées, pouvant contenir des effets de lot, l'ascendance individuelle ou la composition des types cellulaires, sont intégrées dans le modèle de régression en utilisant des facteurs latents. Dans ces modèles, les tailles d'effets de la variable d'intérêt sur les marqueurs génomiques et les facteurs latents sont estimés simultanément.

Les modèles de régression à facteurs latents englobent plusieurs méthodes, incluant l'analyse des variables de substitution (Surrogate Variable Analysis ou SVA, (Leek and Storey, 2007), les modèles mixtes à facteurs latents (Latent Factor Mixed Models ou LFMM, (Frichot et al., 2013)), l'analyse en composantes principales résiduelles (Kalaitzis and Lawrence, 2012), l'analyse directe des variables de substitution (Direct Surrogate Variable Analysis ou dSVA, (Lee et al., 2017)), la régression robuste à rangs réduits (She and Chen, 2017) et les tests et estimations ajustés en fonction de la confusion (Confounder Adjusted Testing and Estimation ou CATE, (Wang et al., 2017)). Chaque

méthode a des mérites spécifiques par rapport à une catégorie d'étude d'association, et les performances des méthodes ont été largement débattues dans des études comparatives (par exemple, voir (Kaushal et al., 2015)).

Les résultats des GWAS et des EWAS sont principalement reportés sous forme de valeurs de significativité (*P*-valeurs) et de tailles d'effet spécifiques à chaque locus. Ce sont des statistiques récapitulatives d'une importance cruciale dans l'interprétation et l'application des résultats (Battram et al., 2021; Buniello et al., 2019). Une propriété commune des modèles de régression à facteurs latents est d'utiliser des paramètres de régularisation induisant des contraintes sur les estimations des tailles d'effets. Parmi ces méthodes, les modèles de régression "parcimonieux" (sparse) supposent qu'une proportion relativement faible de tous les marqueurs génomiques sont corrélés avec la variable d'intérêt, et évaluent les associations en évitant le problème des tests multiples (Hoggart et al., 2008; Tibshirani, 1996; Wu et al., 2009). Pour simplifier la suite de l'étude, tout modèle permettant la sélection de variable via l'inclusion de tailles d'effet non-nulles sera appelé "méthode sparse" tandis que les autres méthodes seront "non-sparse".

Des modèles "sparse" ont également été couplés à des modèles linéaires mixtes, pour combiner les avantages des deux, pour l'étude de traits polygéniques. C'est le cas du modèle bayésien BSLMM (Bayesian Sparse Linear Mixed Model) (Zhou et al., 2013; Zhou and Stephens, 2012).

Dans cette étude, nous introduisons un algorithme des moindres carrés qui estime conjointement les tailles d'effet et les facteurs de confusion dans un modèle de régression “sparse” à facteurs latents. Les tailles d’effets sont estimées grâce à une méthode des moindres carrés, régularisés par la norme L^1 et par une norme nucléaire. Grâce à l'inclusion de contraintes de sparsité, l'algorithme permet d'identifier des tailles d'effet non nulles. Cette approche permet de ne pas utiliser de tests statistiques et donc de passer outre le problème de tests multiples.

Nous appelons notre approche “sparse LFMM” pour “sparse latent factor mixed models”. L’étude est découpée en plusieurs parties. En premier lieu, nous décrivons les modèles à facteurs latents. Ensuite, par des campagnes de simulation, nous comparons sparse LFMM a un ensemble de méthodes de la littérature. Pour finir, nous avons conduit deux analyses de données réelles. D’une part, nous avons utilisé sparse LFMM pour effectuer une GWAS de la période de floraison de la plante *Arabidopsis thaliana* et de l’autre nous avons effectué une EWAS du le tabagisme chez les femmes enceintes.

2.2. Les modèles à facteurs latents (LFMM)

Les modèles de régression à facteurs latents évaluent les associations entre les éléments d'une matrice de réponse (*Y*), et les variables d'intérêt, appelées variables primaires (*X*), mesurées pour *n* individus. La matrice de réponse contient *p* marqueurs moléculaires, qui peuvent représenter tout type de données omiques (génotypes, méthylation de l'ADN, etc.), collectées pour les individus. La matrice *X* peut également incorporer des facteurs de confusion observés tels que l'âge, le sexe, etc., et sa dimension est *n* × *d*, où *d* représente le nombre total de variables. Les termes de modèles de régression à facteurs latents et de modèles mixtes à facteurs latents sont des synonymes. Les modèles mixtes à facteurs latents contiennent des effets fixes et latents, et ces modèles ne doivent pas être confondus avec les modèles linéaires mixtes, qui contiennent des effets fixes et des effets aléatoires. Les modèles de régression des facteurs latents combinent les effets fixes et latents comme il suit (Eq1) :

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{W} + \mathbf{E}.$$

Les tailles d'effet fixes sont enregistrées dans la matrice ***B*** de dimension *p* × *d*. La matrice *E* représente les erreurs résiduelles et a la même dimension que la matrice de réponse. La matrice ***W*** est une matrice latente de rang *K* (Frichot et al., 2013; Lee et al., 2017; Leek and Storey, 2007; Wang et al., 2017). La valeur de *K* est inconnue et doit être estimée. Le terme modèle de régression des facteurs latents se justifie par le fait que la matrice latente W peut être factorisée comme un produit de deux matrices, $W = UV^T$, où la matrice *U*, de dimension *n* × *K*, contient les facteurs latents. Une façon de réaliser la factorisation consiste à utiliser la décomposition en valeurs singulières (SVD) de *W*. Dans

ce cas, *V* est une matrice de charge (loadings) semi-orthogonale de dimension $p \times K$ (Eckart and Young, 1936).

Pour motiver l'utilisation de fonctions de perte pénalisée, nous considérons des estimations statistiques naïves pour les matrices *B* et *W* dans l'équation (1) obtenues par la minimisation d'une fonction de perte des moindres carrés classique (Eq2) :

$$\mathcal{L}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2$$

où $\|.\|_F$ est la norme matricielle de Frobenius. La valeur minimale de la fonction de perte est atteinte lorsque *W* est calculé comme la décomposition en valeurs singulières de rang *K* de *Y*. Dans ce cas, la matrice *B* pourrait être obtenue comme estimation d'une régression linéaire de la matrice résiduelle $(Y - W)$ sur *X*.

Pour justifier l'introduction de termes de régularisation dans la fonction de perte, nous avons remarqué que l'interprétation des composantes principales en tant qu'estimations de facteurs de confusion peut être incorrecte, car elles n'incluent aucune information sur la variable primaire *X*. La construction ci-dessus est donc problématique. Pour le montrer, considérons la décomposition en valeurs singulières de $W = UV^T$. Pour toute matrice *P* de dimensions $d \times p$, nous pouvons vérifier que :

$$\|\mathbf{Y} - (\mathbf{U} - \mathbf{X}\mathbf{P})\mathbf{V}^T + \mathbf{X}(\mathbf{B}^T - \mathbf{P}\mathbf{V}^T))\|_F^2 = \|\mathbf{Y} - \mathbf{U}\mathbf{V}^T + \mathbf{X}\mathbf{B}^T\|_F^2$$

En conséquence, *B* et $(B - VP^T)$ correspondent à des solutions valides du problème de minimisation et sont la preuve qu'il existe un espace infini de solutions. Pour conclure, la fonction de perte doit être modifiée afin de garantir la dépendance de *W* à la fois sur *Y* et *X*, et pour permettre le calcul de solutions bien définies.

2.2.1. Algorithme Sparse LFMM

Pour résoudre les problèmes décrits dans la section ci-dessus, une approche par régularisation a été utilisée. Cette approche introduit des pénalités basées sur la norme L^1 des coefficients de régression et sur la norme nucléaire de la matrice latente :

$$\mathcal{L}_{\text{sparse}}(\mathbf{W}, \mathbf{B}) = \left\|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\right\|_F^2 + \mu\|\mathbf{B}\|_1 + \gamma\|\mathbf{W}\|_*, \quad \mu, \gamma > 0$$

où $\|B\|_1$ désigne la norme L^1 de B, μ est un paramètre de régularisation L^1, W est la matrice latente, $\|W\|_*$ désigne sa norme nucléaire et γ est un paramètre de régularisation pour la norme nucléaire. La pénalité L^1 induit une parcimonie sur les effets fixes (Tibshirani, 1996), et correspond à l'information a priori que toutes les variables de réponse ne peuvent pas être associées aux variables primaires. Plus précisément, l'a priori implique qu'un nombre restreint de lignes de la matrice de tailles d'effet B sont non nulles. Le second terme de régularisation est basé sur la norme nucléaire. Le paramètre de norme nucléaire (γ) détermine le rang de la matrice latente W, et sera choisi pour que le rang soit égal à K. Avec ces termes de pénalité, $\mathcal{L}_{sparse}(W,B)$ est une fonction convexe, et les algorithmes de minimisation convexe peuvent être appliqués pour obtenir des estimations de B et W (Mishra et al., 2013).

Comme (Tibshirani, 1996), nous supposons que les variables primaires, X, sont centrées, de sorte que $X^TX = I$ (I est la matrice identité de dimension $d \times d$) mais le cas général est traité dans le package R *lfmm*. Nous avons développé une méthode de descente par blocs pour minimiser la fonction de perte convexe $\mathcal{L}_{sparse}(W, B)$ par rapport à B et W. L'algorithme est initialisé à partir de la matrice nulle $\widehat{W}_0 = 0$, et itère les étapes suivantes.

1. Trouver $\hat{B}_t$ un minimum de la fonction de perte pénalisée

$$\mathcal{L}^{(1)}_{\text{sparse}}(\mathbf{B}) = \|(\mathbf{Y} - \hat{\mathbf{W}}_{t-1}) - \mathbf{X}\mathbf{B}^T\|^2_F + \mu\|\mathbf{B}\|_1$$

2. Trouver $\widehat{W}_t$ un minimum de la fonction de perte pénalisée

$$\mathcal{L}^{(2)}_{\text{sparse}}(\mathbf{W}) = \|(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^T_t) - \mathbf{W}\|^2_F + \gamma\|\mathbf{W}\|_*$$

L'algorithme parcourt les deux étapes jusqu'à ce qu'un critère de convergence soit satisfait ou que la ressource informatique allouée soit épuisée. Chaque étape de minimisation a une solution bien définie et unique. Pour le voir, notons que l'étape 1 correspond à une régression L^1-régularisée de la matrice résiduelle $(Y - \widehat{W}_{t-1})$ sur les variables explicatives. Pour calculer les coefficients de régression, nous avons utilisé la méthode de descente par blocs de Friedman (Friedman et al., 2007). D'après (Tibshirani, 1996), nous avons obtenu :

$$\hat{\mathbf{B}}_t = \text{sign}(\bar{\mathbf{B}}_t)(\bar{\mathbf{B}}_t - \mu)_+$$

où $s_+ = max(0, s)$, $sign(s)$ est le signe de s, et $\bar{B}_t$ est l'estimation de la régression linéaire, $\bar{B}_t = X^TY - \widehat{W}_{t-1}$. L'étape 2 consiste à trouver une approximation de rang faible de la matrice résiduelle $Y - X\hat{B}_t^T$ (Cai et al., 2008). Le résultat de cette approximation commence par une décomposition en valeurs singulières (SVD) de la matrice résiduelle, $Y - X\hat{B}_t^T = MSN^T$, où M est une matrice unitaire de dimension $n \times n$, N une matrice unitaire de dimension $p \times p$, et S la matrice de valeurs $(s_j)_{j = 1..., n.}$ Ensuite, on obtient :

$$\hat{\mathbf{W}}_t = \mathbf{M}\bar{\mathbf{S}}\mathbf{N}^T$$

où $\bar{S}$ est la matrice diagonale de termes diagonaux $\bar{s}_j = (s_j - \gamma)+, j = 1, . . . , n$.

2.2.2. Algorithme Ridge LFMM

L'algorithme de régression à facteurs latents avec pénalité « ridge » a été décrit pour la première fois dans (Caye et al., 2019). Cette approche est appelée « ridge LFMM ». Les algorithmes sparse LFMM et ridge LFMM sont implémentés dans le même package R et, dans cette partie, nous rappelons comment les estimations des matrices de paramètres *B* et *W* sont calculées pour ridge LFMM. L'approche ridge LFMM minimise la fonction de perte suivante :

$$\mathcal{L}_{\text{ridge}}(\mathbf{B}, \mathbf{W}) = \|\mathbf{Y} - \mathbf{W} - \mathbf{X}\mathbf{B}^T\|_F^2 + \lambda\|\mathbf{B}\|_2^2, \quad \lambda > 0,$$

où W est contrainte à avoir un rang égal à K, $\|.\|_F$ est la norme de Frobenius, $\|.\|_2$ est la norme L^2, et λ est un paramètre de régularisation. L'algorithme de minimisation commence par une SVD de la matrice de variables explicatives, $X = Q\Sigma R^T$, où Q est une matrice unitaire $n \times n$, R est une matrice unitaire $d \times d$ et Σ est une matrice $n \times d$ contenant les valeurs singulières de X, notées par $(\sigma_j)_{j=1,\ldots,d}$. Les estimations de ridge LFMM sont calculées comme il suit :

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y})$$
$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}})$$

où $svd_K(A)$ est la SVD de rang K de la matrice A, I est la matrice identité $d \times d$, et D_λ est la matrice diagonale $n \times n$ avec les coefficients définis comme :

$$\mathbf{d}_\lambda = \left(\sqrt{\frac{\lambda}{\lambda + \sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda + \sigma_d^2}}, 1, \ldots, 1\right)$$

Pour $\lambda > 0$, la solution du problème des moindres carrés régularisés est unique (Caye et al., 2019) et correspond aux équations décrites ci-dessus.

2.2.3. Paramètres de régularisation

Nous avons utilisé plusieurs heuristiques pour choisir les paramètres de régularisation des algorithmes sparse et ridge LFMM. Le paramètre de régularisation L^1, μ, a été déterminé en estimant la proportion de tailles d'effet nulles. Cette proportion a été estimée comme la proportion de *P*-valeurs reflétant H0 à partir de tests statistiques effectués avec ridge LFMM. Elle a été obtenue en utilisant la fonction *pi0est* du package « qvalue » (Storey et al., 2021). Dans ridge LFMM, le paramètre de régularisation a été fixé par défaut à $\lambda = 10^{-5}$ (Caye et al. 2019). Après avoir défini la proportion de tailles d'effet non nulles, μ a été calculé en utilisant l'approche du chemin de régularisation proposée par (Friedman et al., 2010) comme il suit. L'algorithme de chemin de régularisation a été initialisé avec les plus petites valeurs de μ telles que :

$$\mathrm{sign}(\bar{\mathbf{B}}_1)(\bar{\mathbf{B}}_1 - \mu)_+ = 0$$

où $\bar{B}_1$ est l'estimation de la régression linéaire. Ensuite, nous avons construit une suite de valeurs μ qui décroit de la valeur déduite du paramètre μ^{max} à $\mu^{min} = \varepsilon\mu^{max}$. Nous avons finalement mesuré le nombre d'éléments non nuls dans $\hat{B}_t$, et nous nous arrêtons lorsque la proportion cible, 1 - *pi0est,* est atteinte. Nous avons utilisé une approche heuristique pour évaluer γ à partir du nombre de facteurs latents *K*. A partir des valeurs singulières $(\lambda_1, \ldots, \lambda_n)$ de la matrice D_0Q^TY, nous posons :

$$\gamma = \frac{(\lambda_K + \lambda_{K+1})}{2}$$

Avec cette valeur de γ, sparse LFMM converge vers des estimations de matrice de facteurs latents dont le rang est égal à K.

2.3. Évaluation de Sparse LFMM sur des données simulées

Suite au développement de Sparse LFMM, il a été nécessaire de tester son fonctionnement ainsi qu'évaluer ces performances contre un ensemble de méthodes. Nous avons utilisé deux campagnes de simulation, l'une sur la base d'un simulateur de données utilisant le modèle génératif de LFMM (équation 1) et l'autre sur la base d'un simulateur utilisant des données réelles. Notre méthode a été comparée à deux méthodes dites "sparse" (LASSO et BSLMM) ainsi qu'à trois méthodes dites "non-sparse" (SVA, Ridge LFMM et CATE).

Comme référence de méthode "sparse", nous avons utilisé les modèles de régression LASSO (Least Absolute Shrinkage and Selection Operator) (Friedman et al., 2010; Tibshirani, 1996). La régression LASSO n'inclut aucune correction pour la confusion, et induit de forts biais sur les estimations des tailles d'effets. Cette méthode est implémentée dans le package R *glmnet*, et le paramètre de régularisation a été sélectionné en utilisant une approche de validation croisée (5-fold) (Zeng et al., 2017).

Nous avons également utilisé BSLMM (Bayesian Sparse Linear Mixed Models) implémentés dans le logiciel GEMMA (Zhou et al., 2013). BSLMM est une méthode hybride qui combine des modèles de régression "sparses" avec des modèles linéaires mixtes. BSLMM utilise une méthode de Monte-Carlo par chaînes de Markov (MCMC) pour estimer les tailles d'effet. Les paramètres de l'algorithme MCMC ont été fixés à 10 000 cycles (burn-in et sampling) (Zeng et al., 2017). Pour déterminer la proportion de tailles d'effet non nulles dans BSLMM, deux paramètres ont été ajustés (*p-min* et *p-max*).

Ces paramètres correspondent au logarithme des proportions maximales et minimales attendues pour les tailles d'effet non nulles. Notons que BSLMM a uniquement été développé pour réaliser des GWAS et ne peut pas être utilisé pour réaliser des EWAS, car la matrice de réponse est limitée à une matrice de génotypes.

Pour comparer sparse LFMM à des méthodes “non-sparse”, nous avons utilisé la méthode d’analyse par variables de substitution (SVA, (Leek and Storey, 2007)) qui est implémenté dans le package R sva. Ensuite, nous avons utilisé la méthode Confounder Adjusted Testing and Estimation (CATE) (Wang et al., 2017). CATE utilise une transformation linéaire de la matrice de réponse telle que le premier axe de cette transformation est colinéaire à X et les autres axes sont orthogonaux à X. CATE est implémenté dans le package R cate, et a été utilisé sans contrôle négatif. Finalement nous avons utilisé ridge LFMM, un algorithme du package R LFMM (Caye et al., 2019).

2.3.1. Simulation de données de méthylation d'ADN sur la base d'un modèle génératif

Dans une première série d'expériences de type EWAS, nous avons comparé l'algorithme sparse LFMM avec LASSO et trois approches non-sparse (ridge LFMM, CATE, SVA). La performance de chaque méthode a été mesurée dans différents scénarios : tailles d'effet élevées, moyennes ou faibles et en faisant varier les intensités de confusion.

2.3.1.1. Simulateur de données

Nous avons simulé une variable primaire ($d = 1$), $K = 6$ facteurs latents et une matrice réponse selon un modèle gaussien multivarié. La distribution jointe de (U, X) était N (0, S), où S a des termes diagonaux ($s^2_1, \ldots, s^2_K, 1$) et des termes non diagonaux mis à zéro, à l'exception de la covariance entre U_k et X, qui a été définie par $c_k\rho$. Les coefficients c_k ont été échantillonnés à partir d'une distribution uniforme et ρ était proportionnel à $\frac{1}{\sqrt{\sum_k \frac{c_k^2}{s_k^2}}}$. Les écarts types ($s_k$) ont été échantillonnés au hasard dans l'intervalle (2, 6) pour $k < 5$ et dans l'intervalle (0, 1) pour $k = 5$ et 6. Le coefficient de proportionnalité a été choisi de manière à ce que l'intensité de confusion puisse prendre des valeurs relativement faibles ($R = 0.1$) ou élevées ($R = 0.5$).

Pour créer des simulations "sparses", seule une petite proportion de tailles d'effet, égale à 0.8 %, est différente de zéro. Des tailles d'effet non nulles ont été échantillonnées selon une distribution gaussienne, $N(B, 0.2)$, où B pouvait prendre trois valeurs, $B = 0.75$, (low value), $B = 1.5$ (medium value) et $B = 3.0$ (high value). Les erreurs résiduelles et les

charges *V*, ont été échantillonnées selon une distribution gaussienne standard. Les dimensions de la matrice de réponse ont été fixées à *n = 400* individus et *p = 10000* marqueurs. La matrice de réponse a finalement été créée en simulant à partir du modèle génératif suivant, et en utilisant un lien *probit* afin d'imiter les données de méthylation de l'ADN :

$$M_{n,p} = a_{1,p} * X_{n,1} + U_{n,k} * V_{k,p} + E$$

Deux cents simulations ont été réalisées pour chaque combinaison de paramètres. Pour chaque simulation et chaque méthode, nous avons rapporté les erreurs statistiques (RMSE) des estimations des tailles d'effet. Pour fournir une valeur de référence de RMSE, nous avons mesuré l'erreur lorsque toutes les tailles d'effet sont estimées comme étant nulles (valeur « zéro » ou erreur de modèle nul). Nous avons aussi calculé l'indice RMSE uniquement sur l'ensemble des marqueurs causaux. De plus, pour évaluer les capacités des méthodes à identifier les vrais positifs, nous avons utilisé la précision, qui correspond à la proportion de vrais positifs dans une liste de marqueurs positifs, le rappel, qui est le nombre de vrais positifs divisé par le nombre de marqueurs causaux, et le F1-score, qui est la moyenne harmonique de la précision et du rappel.

$$F1 - score = 2\,\frac{precision\,.rappel}{precision + rappel}$$

$$precision = \frac{TP}{TP + FP}$$

$$rappel = \frac{TP}{TP + FN}$$

où TP correspond au nombre de vrais positifs, FP aux faux positifs et FN aux faux négatifs.

Pour calculer la précision et le F1-score dans les simulations de modèles génératifs, une liste de *N = 100* marqueurs ayant les plus grandes tailles d'effet estimées en valeurs absolues a été considérée pour chaque méthode. La longueur de la liste, N = 100, a été choisie comme étant représentative du nombre de marqueurs pouvant être soumis à une validation expérimentale. Pour cette valeur, la précision et le F1-score ne peuvent pas être supérieurs à 80 % et 89 % respectivement.

2.3.1.2. Résultats

Les différents algorithmes ont été utilisés avec leurs paramètres par défaut. Le nombre de facteurs latents a été fixé à *K = 6* dans les méthodes à facteurs latents, et la proportion d'hypothèse nulle a été fixée à 1% pour les méthodes “sparse”. L'erreur du modèle nul est égal à 0.069 dans les scénarios de faibles tailles d'effet et égal à 0.135 dans les scénarios à fortes tailles d'effet. Les RMSE de sparse LFMM varient de 0.055 à 0.092. Les RMSE de LASSO sont proches de ceux de sparse LFMM pour les scénarios ayant de faibles tailles d’effet.

En revanche, les méthodes “non-sparse” ont conduit à des RMSE plus élevés que le modèle-nul, allant entre 0.13 et 0.26 pour ridge LFMM et CATE, et jusqu'à 0.50 pour SVA. Pour les tailles d'effet des marqueurs causaux, les méthodes “non-sparse” ont atteint des RMSE inférieurs à celles des méthodes “sparse”, entre 0.12 et 0.26 pour ridge LFMM et CATE, et entre 0.60 et 1.03 pour sparse LFMM (Figures 2.1 et 2.2).

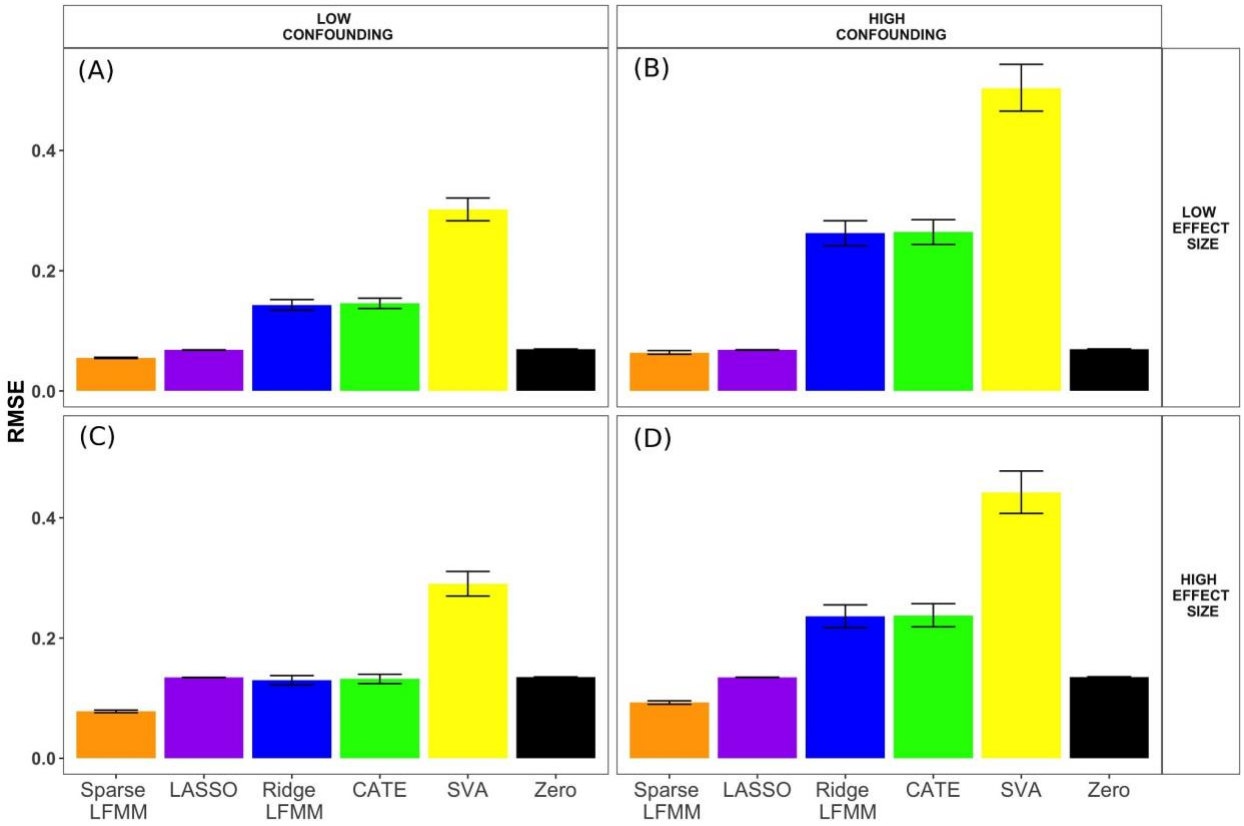

Figure 2.1 : Root Mean Square Error (RMSE) en fonction des tailles d'effet des marqueurs causaux et de l'intensité de confusion. *Deux méthodes "sparse" (sparse LFMM, LASSO) et trois méthodes "non-sparse" (ridge LFMM, CATE et SVA) ont été comparées pour différents niveaux de taille d'effet et d'intensité de confusion.*

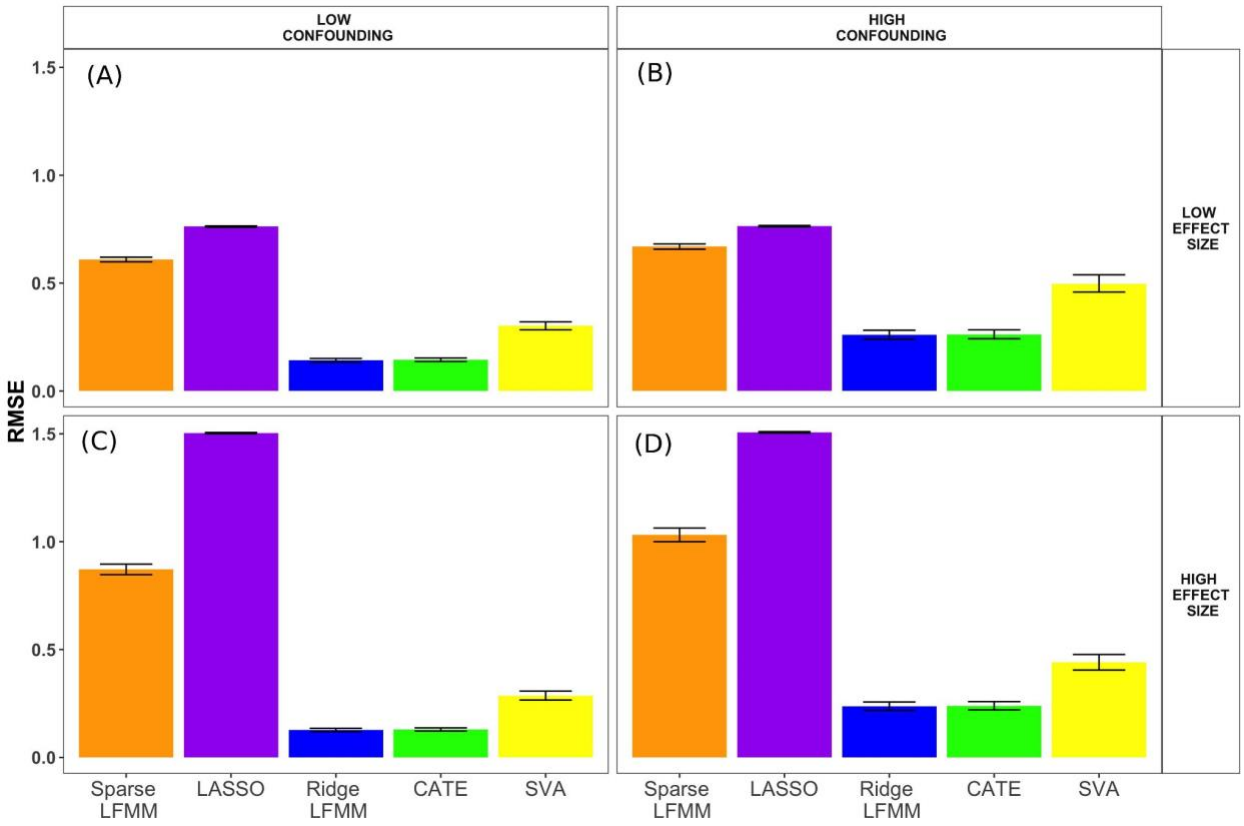

Figure 2.2 : Root Mean Square Error (RMSE) uniquement sur les marqueurs causaux en fonction des tailles d'effet des marqueurs causaux et de l'intensité de confusion. *Deux méthodes "sparse" (sparse LFMM, LASSO) et trois méthodes "non-sparse" (ridge LFMM, CATE et SVA) ont été comparées pour différents niveaux de taille d'effet et d'intensité de confusion.*

Pour toutes les méthodes, la précision et le F1-score sont plus élevés dans les scénarios avec de fortes tailles d'effet et une intensité de confusion plus faible. Dans tous les scénarios, sparse LFMM a obtenu de meilleurs scores que SVA. Sparse LFMM atteint des performances supérieures au LASSO dans les scénarios ayant de fortes (ou moyennes) tailles d'effets (Figure 2.3). Dans les scénarios de forte intensité de confusion sparse LFMM obtient des performances plus faibles que ridge LFMM et CATE. La différence de performance est importante lorsque les tailles d'effets sont faibles (F1-score ≈ 0.68, contre F1-score ≈ 0.85), mais ces différences sont minimes pour les autres

scénarios. Les méthodes obtiennent leurs meilleures performances dans les scénarios ayant de fortes tailles d'effet : sparse LFMM, ridge LFMM et CATE sont proches du F1-score maximal (0.89) et de la précision maximale (0.80). Les AUC sont très fortement corrélées aux F1-scores, et ne modifient pas l'ordre de performance des méthodes.

En résumé, sparse LFMM est généralement préférable au LASSO et à SVA dans les simulations basées sur le modèle génératif et sparse LFMM est associé à la plus petite erreur statistique sur l''ensemble des ensimations.

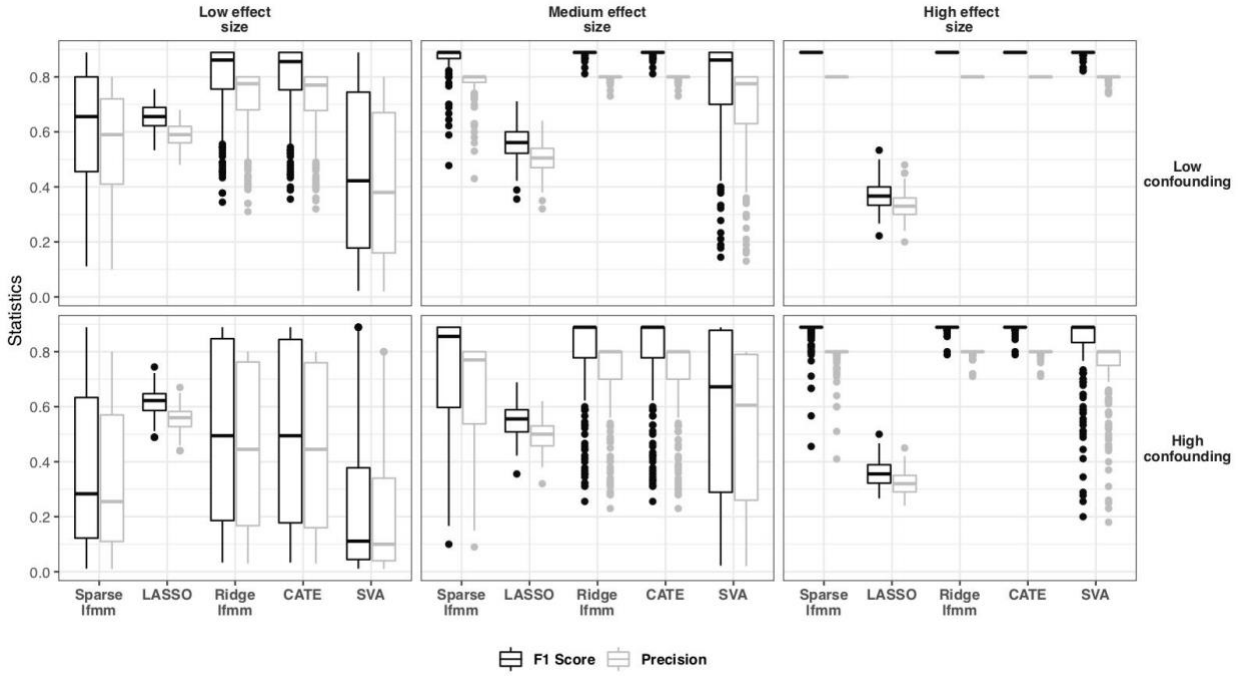

Figure 2.3 : F1-score et précision des méthodes "sparse" et "non-sparse" dans les simulations à partir du modèle génératif. *Deux méthodes "sparse" (sparse LFMM, LASSO) et trois méthodes "non-sparse" (ridge LFMM, CATE et SVA) ont été comparées pour différents niveaux de taille d'effet et d'intensité de confusion. Le F1-score est défini comme la moyenne harmonique de la précision et du rappel.*

2.3.2. Simulation de données génotypiques à partir de données réelles

Dans une deuxième série d'expériences, nous avons utilisé des simulations réalistes pour évaluer et comparer sparse LFMM aux autres approches. Ces simulations ne sont pas réalisées à partir d'un modèle génératif mais à partir de données génétiques réelles.

2.3.2.1. Simulateur de données

Des phénotypes ont été simulés pour $n = 162$ individus à partir de données publiques de polymorphisme d'un seul nucléotide (SNP) du chromosome 5 de la plante modèle *Arabidopsis thaliana* (Atwell et al., 2010). La matrice de réponse contient $p = 53859$ SNPs. Les simulations de phénotypes incorporent des caractéristiques réalistes telles que des interactions gène-environnement (GxE) à l'aide de variables environnementales extraites d'une base de données bioclimatiques (François and Caye, 2018).

Les simulations ont été effectuées comme suit. Considérant un sous-ensemble prédéfini de marqueurs causaux (J), un trait phénotypique, x_i, a été créé pour chaque individu à partir du modèle suivant :

$$x_i = \sum_{j \in J} \beta_j y_{ij} + \sum_{j \in J} \delta_j x_{ij} e_i + \sum_{k=1}^{K} u_{ik} + \epsilon_i$$

Où y_{ij} représente le génotype de l'individu i au locus j, e_i correspond à la première composante principale de 19 variables extraites de la base de données bioclimatiques, u_k représente les facteurs de confusion, et ϵ_i est une matrice d'erreur résiduelle. Les paramètres β (ou B) représentent les tailles d'effet des marqueurs causaux. Le nombre de facteurs de confusion est $K = 6$, et les phénotypes ont été générés à partir de cinq

marqueurs causaux ayant des effets identiques. Deux valeurs de taille d'effet ont été utilisées, B = 6 (faible taille d'effet) et B = 9 (forte taille d'effet). Les interactions GxE sont représentées par le paramètre δ. Deux valeurs d'interaction gène par environnement ont été considérées, GxE = 0.1 (faible GxE) et GxE = 0.9 (fort GxE). Pour chaque combinaison de paramètres, 200 simulations ont été réalisées. Dans ces simulations empiriques, le F1-score a été modifié pour tenir compte du déséquilibre de liaison (LD) dans les données : les marqueurs présent dans une fenêtre de taille de 10kb autour d'un marqueur causal ont été considérés comme de vraies découvertes (LD-r^2 < 0.2, (François and Caye, 2018)).

2.3.2.2. Résultats

Dans les scénarios de faible intensité GxE, sparse LFMM a obtenu les F1-scores les plus élevés (F1-score entre 0,57 à 0,60, précision entre 0,81 et 0,82, figure 2.4) par rapport à BSLMM (F1 score entre 0,36 et 0,44) et aux méthodes “non-sparse” (F1-score entre 0,25 et 0,28). Dans les scénarios de forts GxE, toutes les méthodes obtiennent de mauvaises performances pour de faibles tailles d'effet, mais sparse LFMM fait partie des méthodes qui obtiennent les meilleurs F1-score et précision. Lorsque les tailles d’effet sont plus élevées, sparse LFMM atteint de meilleures performances (F1-score de 0,28 et précision de 0,33) que les autres méthodes (Figure 2.4D).

Dans ces simulations réalistes, sparse LFMM démontre une plus grande robustesse que les méthodes sparses BSLMM et LASSO, et se compare favorablement aux méthodes “non-sparse” (ridge LFMM, CATE et SVA).

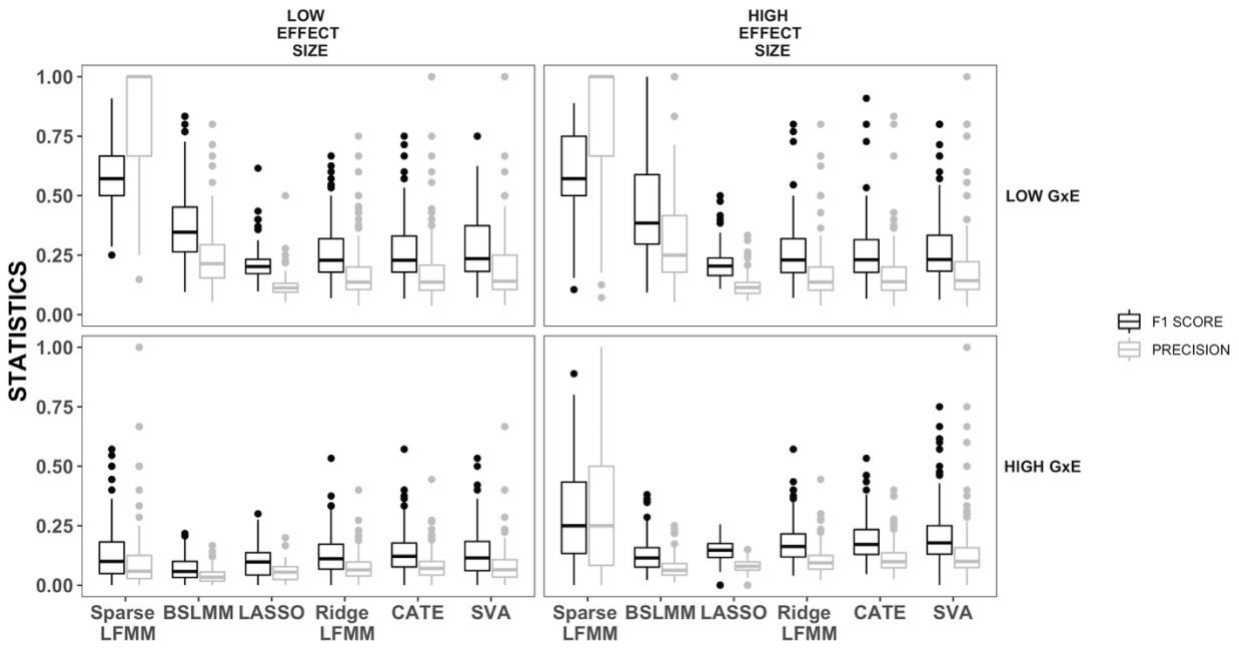

Figure 2.4 : F1-score et précision des méthodes "sparse" et "non-sparse" dans les simulations empiriques. *Trois méthodes "sparse" (sparse LFMM, LASSO et BSLMM) et trois méthodes "non-sparse" (ridge LFMM, CATE et SVA) ont été comparées pour différents niveaux de tailles d'effet et d'interaction GxE. Le F1-score est défini comme la moyenne harmonique de la précision et du rappel.*

2.3.2.3. Temps de calcul

Pour finir avec les simulations, nous avons évalué les temps d'exécution de sparse LFMM (Figure 2.5). Quel que soit le nombre d'individus ou de marqueurs, ridge LFMM est la méthode la plus rapide tandis que sparse LFMM est la plus lente. Ceci est facilement explicable car sparse LFMM itère de nombreux cycles avant de converger vers un résultat, alors que ridge LFMM est une approche exacte. Il a fallu environ 2000 secondes (≈ 33 minutes) pour que sparse LFMM produise un résultat avec n = 1000 individus et p = 100000 marqueurs, alors que ridge LFMM a pris quelques secondes pour calculer les estimations. Étant donné que le temps d'obtention des données de n'importe quelle

analyse peut facilement se compter en années, quelques minutes de temps de calcul est une durée tout à fait acceptable.

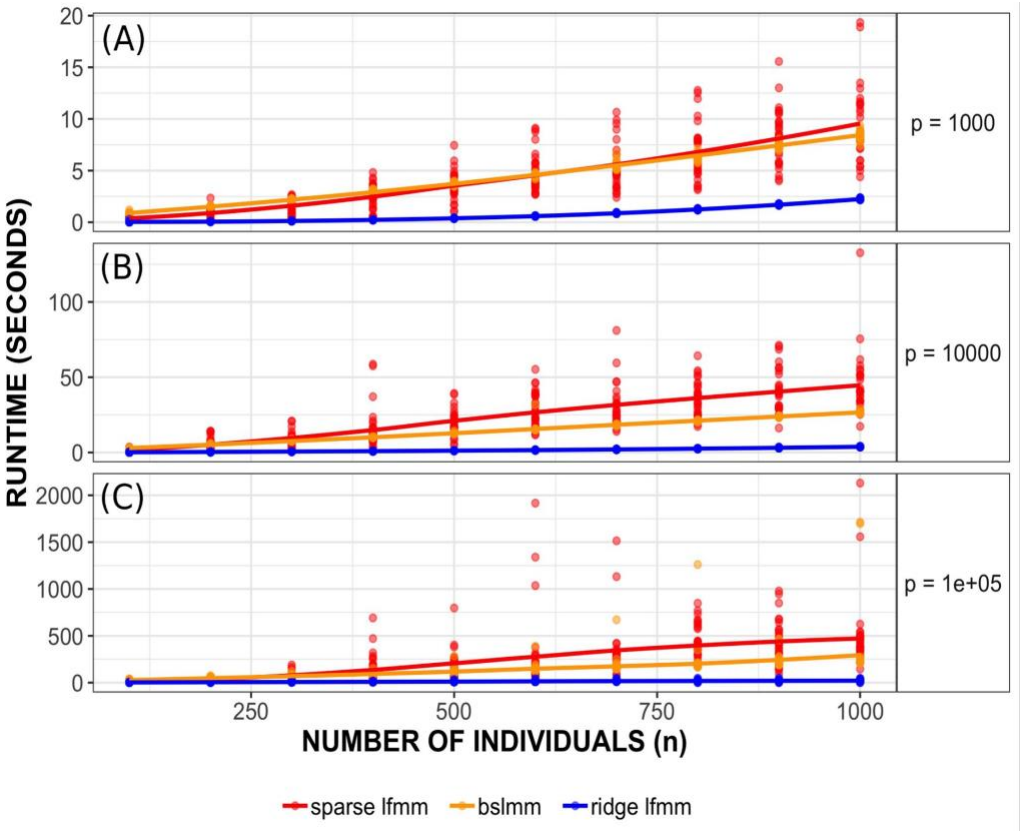

Figure 2.5 : Comparaison des temps d'exécution de trois méthodes. *Temps de calcul en fonction du nombre de marqueurs (p) et du nombre d'individus (n). (A) p = 1000. (B) p = 10 000. (C) p = 100 000.*

2.3.3. Résumé des simulations

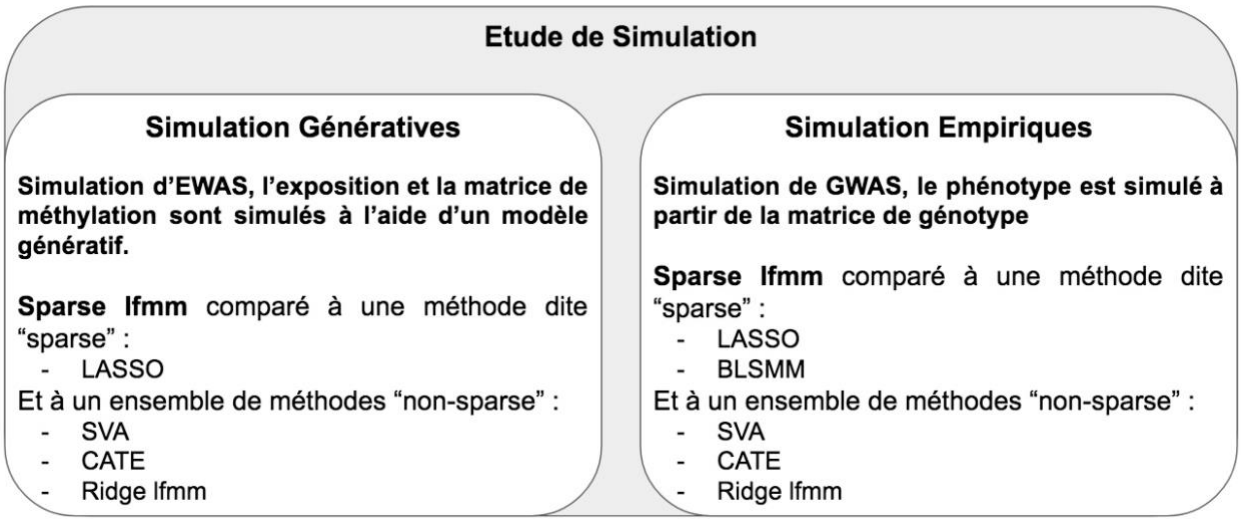

Figure 2.6 : Résumé des études de simulation réalisé pour évaluer sparse LFMM.

Dans les simulations génératives et empiriques (Figure 2.6), sparse LFMM a obtenu des F1-scores et des précisions supérieurs aux méthodes “sparse”, précédemment introduites, BSLMM et LASSO. Par rapport aux trois méthodes “non-sparse” (ridge LFMM, CATE et SVA), les RMSE des estimations de tailles d'effet sur l’ensemble des marqueurs sont globalement plus faibles chez sparse LFMM. Dans les simulations génératives, ridge LFMM et CATE ont atteint des précisions et des F1-scores plus élevés que sparse LFMM. Une explication de ce résultat peut être que les simulations génératives favorisent les méthodes pour lesquelles l'estimation des facteurs latents n'est pas contrainte, et donc la pénalité de norme nucléaire appliquée à la matrice latente dans sparse LFMM peut augmenter le biais dans les estimations des tailles d'effet. En revanche, sparse LFMM obtient ces performances les plus élevées dans les simulations empiriques.

En étant plus robustes à l'écart des hypothèses du modèle (un faible nombre de marqueurs ayant un effet sur le phénotype), les simulations basées sur des données réelles ont montré que sparse LFMM peut surpasser les méthodes “non-sparse”. (Lotterhos, 2019) aboutit à une conclusion similaire basée sur des simulations en génétique des populations. Une limitation de sparse LFMM est son temps d'exécution relativement lent, empêchant une exploration approfondie des choix des hyperparamètres dans les simulations.

2.4. Utilisation de Sparse LFMM sur des données réelles

Dans l'optique d'illustrer notre méthode, nous avons conduit deux analyses de données réelles. Pour démontrer le caractère polyvalent de notre méthode, la première analyse est une GWAS tandis que la seconde est une EWAS. La première analyse consiste à expliquer un phénotype de floraison chez *Arabidopsis thaliana* par des mesures polymorphismes nucléotidiques (SNP) et la seconde vise à comprendre l'impact du tabagisme maternel sur la méthylation de l'ADN placentaire.

2.4.1. Étude d'association à l'échelle du génome d'un phénotype de floraison chez *Arabidopsis thaliana*

Pour notre première analyse, les données portent sur la plante modèle *Arabidopsis thaliana*. Nous avons considéré $n = 162$ individus européens et $p = 53859$ *SNPs* du chromosome 5. Nous avons étudié les associations entre les données de SNP et un phénotype de floraison (FT16-TO : 0000344, (Atwell et al., 2010)). FT16 est un phénotype qui correspond au nombre de jours nécessaires à une plante individuelle pour atteindre le stade de floraison.

2.4.1.1. Méthodes

Dans sparse LFMM, le pourcentage de taille d'effet non nulle a été fixé à 1 %, en accord avec la proportion d'hypothèse nulle estimée à partir des approches "non-sparse". Pour toutes les méthodes à facteurs latents (sparse LFMM, SVA, CATE et ridge LFMM), le nombre de facteurs latents a été fixé à $K = 10$, déterminé par l'observation de l'éboulis des valeurs propres d'une analyse en composante principale réalisée sur les données de

SNP. Les paramètres *pmin* et *pmax* définissant la “sparsité” (ou parcimonie) dans l'algorithme BSLMM ont été fixés à *pmin = -5* et *pmax = -4* respectivement. Ces valeurs correspondent au logarithme des proportions attendues de tailles d'effet non nulles.

2.4.1.2. Résultats

Les méthodes “sparse” (sparse LFMM, LASSO, BSLMM) diffèrent dans leur estimation de la proportion de tailles d'effet nulles (Figure 2.7). LASSO a estimé 99,85 % de tailles d’effet nulles alors que les proportions sont respectivement égales à 99,24 % et 98,18 % pour BSLMM et sparse LFMM. LASSO est l'approche la plus conservatrice, et sparse LFMM la plus libérale. Sparse LFMM partage 3,9% des hits avec LASSO et 5,5% avec BSLMM (Figure 2.8). Moins de 1 % de tous les résultats étaient communs aux trois approches. Les tailles d'effet (non nulles) pour les hits varient sur des échelles distinctes, LASSO présentant les biais les plus importants. Toutes les méthodes “sparse” ont détecté le même top hit à environ 4 Mb, correspondant à un SNP situé dans le gène FLC, qui est cohérent avec les résultats de (Atwell et al., 2010). Le deuxième hit selon Atwell et al., est situé dans le gène DOG1, et il est également identifié par sparse LFMM. BSLMM a plus de difficultés à identifier les deux gènes découverts précédemment.

Compte tenu de la forte corrélation (supérieure à 94 %) entre les tailles d'effet obtenues par les méthodes “non-sparse”, nous avons regroupé leurs résultats en faisant la moyenne de leurs estimations. Les méthodes “non-sparse” ont montré des tailles d'effet dans une gamme de valeurs plus proches de sparse LFMM que de LASSO et de BSLMM, mais ces méthodes obtiennent des erreurs statistiques plus élevées (Figure 2.7). Dans l'ensemble, nous avons trouvé une corrélation significative entre les tailles d'effet non

nulles estimées par sparse LFMM et les tailles d'effet correspondantes trouvées par des méthodes “non-sparse” ($r = 0,8065$, $P < 10^{-16}$). Pour finir, sparse LFMM et les méthodes “non-sparse” ont détecté de nouveaux hits autour de 13,9 Mb et 6,5 Mb, correspondant respectivement aux gènes SAP et ACL5.

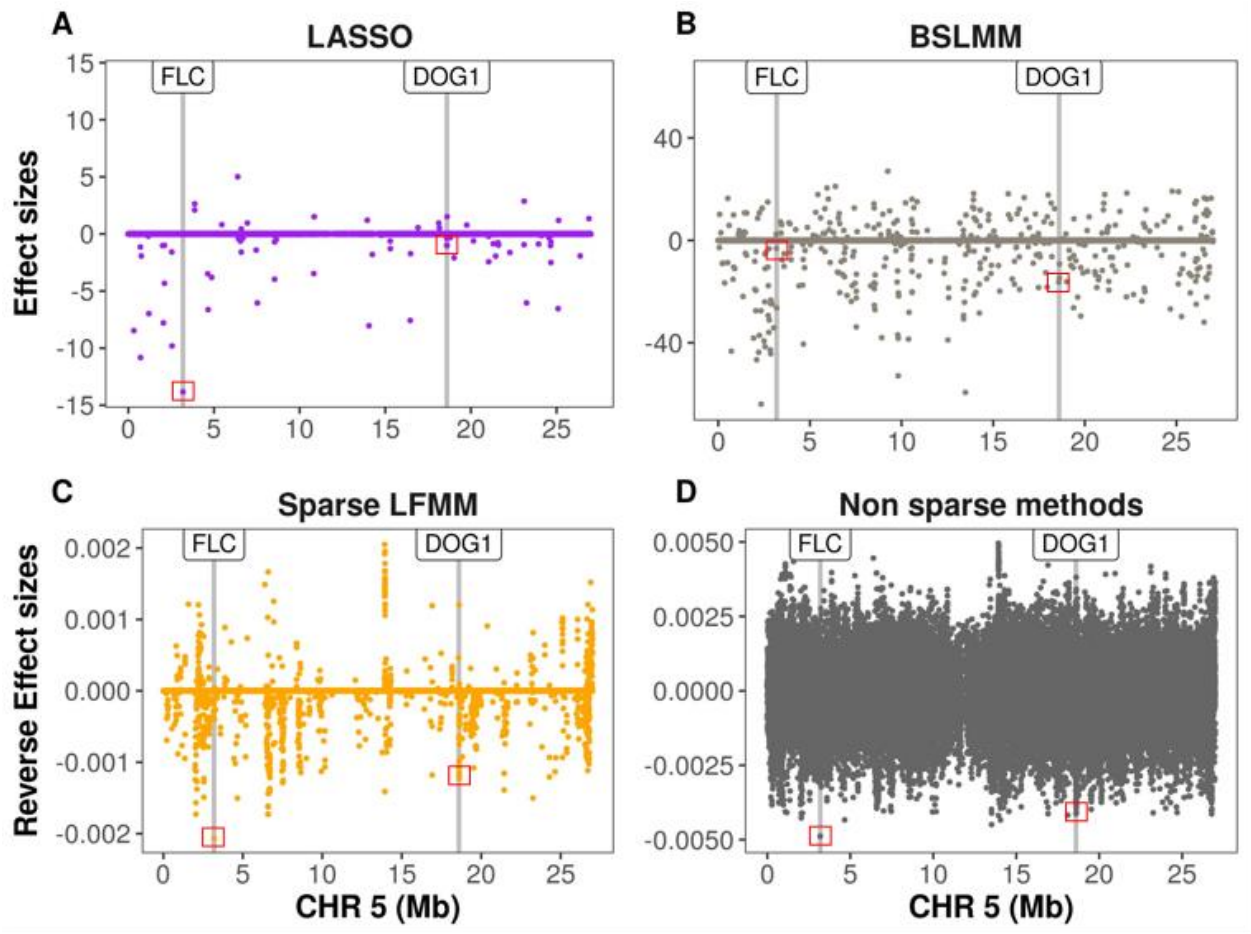

Figure 2.7 : GWAS sur le phénotype de floraison FT16 chez Arabidopsis thaliana (chromosome 5). *A) Estimations des tailles d'effet pour LASSO. B) Estimations des tailles d'effet pour BSLMM. C) Estimations des tailles d'effet inverse pour sparse LFMM. D) Moyenne des estimations des tailles d'effet inverse pour les méthodes “non-sparse” (ridge LFMM, CATE et SVA). Les barres grises représentent les SNP d'Arabidopsis associés au phénotype FT16 dans (Atwell et al., 2010), et correspondent aux gènes FLC et DOG1.*

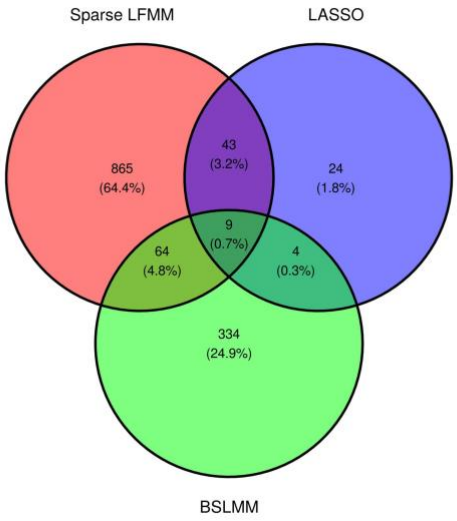

Figure 2.8 : GWAS sur le phénotype de floraison FT16 chez Arabidopsis thaliana par les méthodes "sparse". *Diagramme de Venn des hits associés au phénotype FT16 pour chaque approche. Les hits correspondent à des SNP ayant des estimations de tailles d'effet non nulles.*

2.4.1.3. Discussion

Pour démontrer l'application de notre méthode sur des données de SNP nous avons donc effectué une GWAS. Elle portait sur la floraison (phénotype FT16) d'*Arabidopsis thaliana* chez 162 individus européens.

Sparse LFMM a identifié les gènes FLC et DOG1 associés au phénotype FT16. Les deux gènes ont été précédemment rapportés comme étant associés à FT16 dans (Atwell et al., 2010). Le gène FLC joue un rôle central dans la floraison induite par la vernalisation (Sheldon et al., 2000), et DOG1 est impliqué dans le contrôle de la dormance et de la germination des graines (Nishimura et al., 2018). De plus, sparse LFMM a permis de découvrir de nouveaux hits, comme les SNPs situés sur le gène SAP, un régulateur transcriptionnel impliqué dans la spécification de l'identité florale (Byzova et al., 1999).

Cette association a également été retrouvée dans les méthodes “non-sparse”. De plus, sparse LFMM a détecté des SNPs situés dans le gène ACL5, qui joue un rôle dans la croissance internodale et la taille des organes (Hanzawa et al., 1997). En résumé, les méthodes “sparse” facilitent la sélection des SNPs via leurs tailles d'effet non nulles et ce sans avoir recours à des tests statistiques. Alors que pour les méthodes “non-sparse”, pour réaliser de la sélection de variables, il est nécessaire de réaliser des tests statistiques et d’appliquer des seuils de significativité, sachant ces tests statistiques induisent le problème des tests multiples. Les résultats de sparse LFMM sont non seulement cohérents avec les découvertes précédentes, mais ils ont également permis d'identifier de nouveaux gènes liés au phénotype de floraison FT16 ayant des annotations fonctionnelles intéressantes.

2.4.2. Étude de l’impact du tabagisme maternel sur la méthylation de l’ADN placentaire

Concernant notre seconde analyse de données, nous avons réalisé une EWAS pour estimer les effets du tabagisme maternel sur la méthylation placentaire.

2.4.2.1. Méthodes

Nous avons considéré les niveaux de méthylation bêta-normalisés pour *p = 425 878* CpGs et *n = 668* femmes. Nous avons testé l'association entre le statut tabagique (219 femmes fumeuses et 449 femmes non fumeuses) et les niveaux de méthylation de l'ADN (ADNm) dans les tissus placentaires. Ces données nous viennent de la cohorte mère-enfant EDEN (Heude et al., 2016). Pour la proportion de tailles d'effet non nulles dans sparse LFMM, une valeur conservatrice de 0,1 % a été utilisée (Rousseaux et al., 2020). Pour les modèles à facteurs latents, le nombre de facteurs latents a été fixé à *K = 7*.

2.4.2.2. Résultats

Sparse LFMM retrouve une proportion de tailles d'effet nulles égale à 99,698%, équivalente à 1287 hits (Figure 2.9). Pour caractériser ces 1287 CpG, nous avons évalué s'il y avait un enrichissement en régions amplificatrices (enhancers) et promotrices par rapport à l’ensemble du méthylome (Figure 2.10 et Figure 2.11). Les régions promotrices et amplificatrices ont été identifiées selon les annotations de la puce Illumina (HumanMethylation450 BeadChip). Une proportion de 25,48 % de hits a été retrouvée dans des régions amplificatrices, contre 22,73 % sur l’ensemble du méthylome, et 6,83 % ont été trouvées dans des régions promotrices, contre 19,94 % pour le méthylome.

Ensuite, nous avons comparé les CpGs ayant les tailles d'effet les plus élevées par chaque méthode (Figure 2.12). Sparse LFMM partage 45,3 % de ses hits avec les modèles non-sparses (représentés par le Ridge LFMM) et 2,8 % de ses hits avec LASSO. Comme pour l'étude de GWAS, c'est la moyenne des tailles d'effet des méthodes "non-sparse" qui est montrée dans la figure car les corrélations entre ces méthodes sont supérieures à 99%. Parmi les 51 top hits partagés entre sparse LFMM et ridge LFMM, 25 sont dans le corps d'un gène, 11 ne sont pas associés à un gène, 20 sont situés dans des régions amplificatrices et 2 dans des régions promotrices. Les résultats de sparse LFMM concordent mieux avec les résultats des méthodes "non-sparse" qu'avec ceux de LASSO. La corrélation entre les tailles d'effet non nulles estimées par sparse LFMM et les tailles d'effet correspondantes estimées par les méthodes "non-sparse" est égale à 80,38% ($P < 10^{-16}$), tandis que la corrélation entre les tailles d'effet non nulles de sparse LFMM et de LASSO est égale à 61,86 % ($P < 10^{-16}$).

Pour détailler les résultats dans une région génomique spécifique, nous nous sommes concentrés sur le chromosome 3, qui contient le top hit à l'échelle de l'épigénome pour sparse LFMM, pour les méthodes "non-sparse" (cg27402634, situé dans une région amplificatrice, figure 2.9), et pour LASSO. Les hits de sparse LFMM partagent trois CpGs avec des méthodes "non-sparse" : cg09627057, cg18557837 et cg12662091. Sur l'ensemble du chromosome 3, sparse LFMM a détecté 61 CpGs avec des tailles d'effet non nulles : 43 sont situés dans les gènes, 22 dans les régions amplificatrices et 6 dans les régions promotrices.

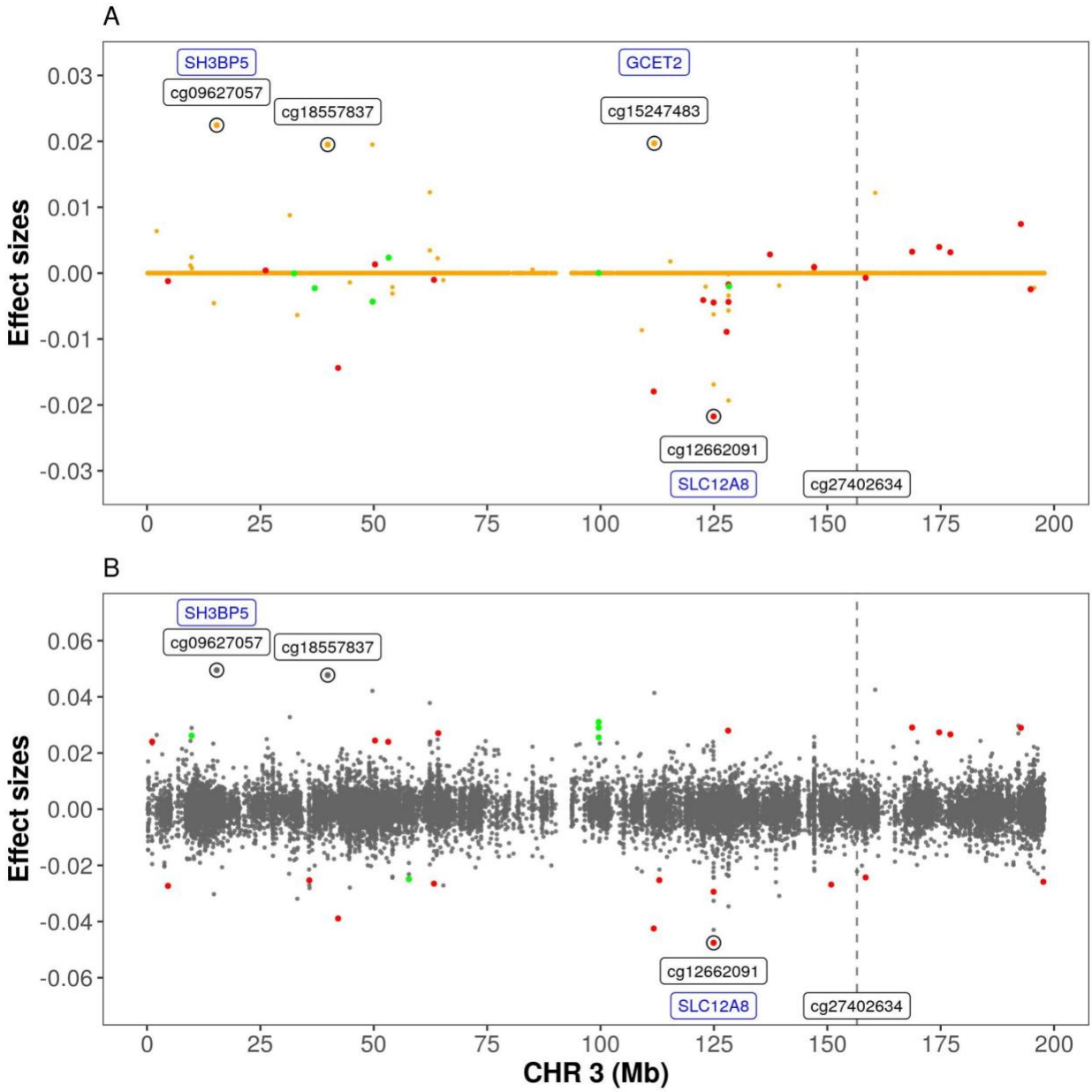

Figure 2.9 : EWAS sur le statut tabagique chez les femmes enceintes (chromosome 3). *A) Taille d'effet estimée pour sparse LFMM. La taille de l'effet de cg27402634 est égale à = -0,117 (hors limites). B) Taille d'effet estimée pour les méthodes "non-sparse" (ridge LFMM, CATE et SVA). La taille de l'effet de cg27402634 est égale à = -0,141 (hors limites). Les CpG ayant les tailles d'effet les plus élevées sont encerclés (gènes en bleu). Les points rouges représentent les CpGs situés dans les régions enhancers. Les points verts représentent les CpGs situés dans les régions promotrices.*

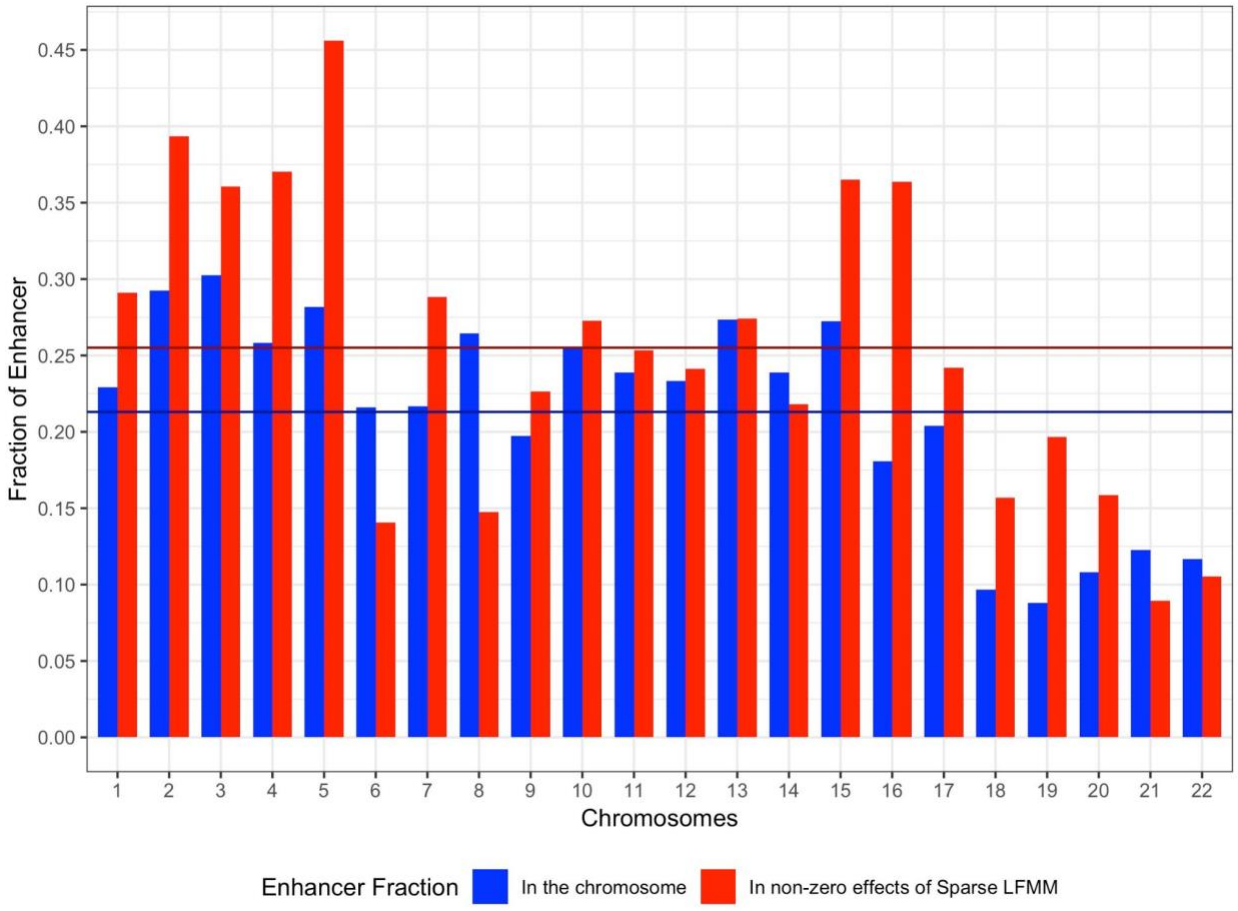

Figure 2.10 : EWAS sur le statut tabagique chez les femmes enceintes. *Surreprésentation des régions amplificatrices dans les hits de sparse LFMM par rapport à l'ensemble du méthylome. Les barres bleues correspondent à la proportion de régions amplificatrices dans chaque chromosome. Les barres rouges correspondent à la proportion de régions amplificatrices détectées par sparse LFMM. La ligne bleue horizontale représente la proportion moyenne de régions amplificatrices pour le méthylome. La ligne rouge représente la proportion moyenne pour sparse LFMM.*

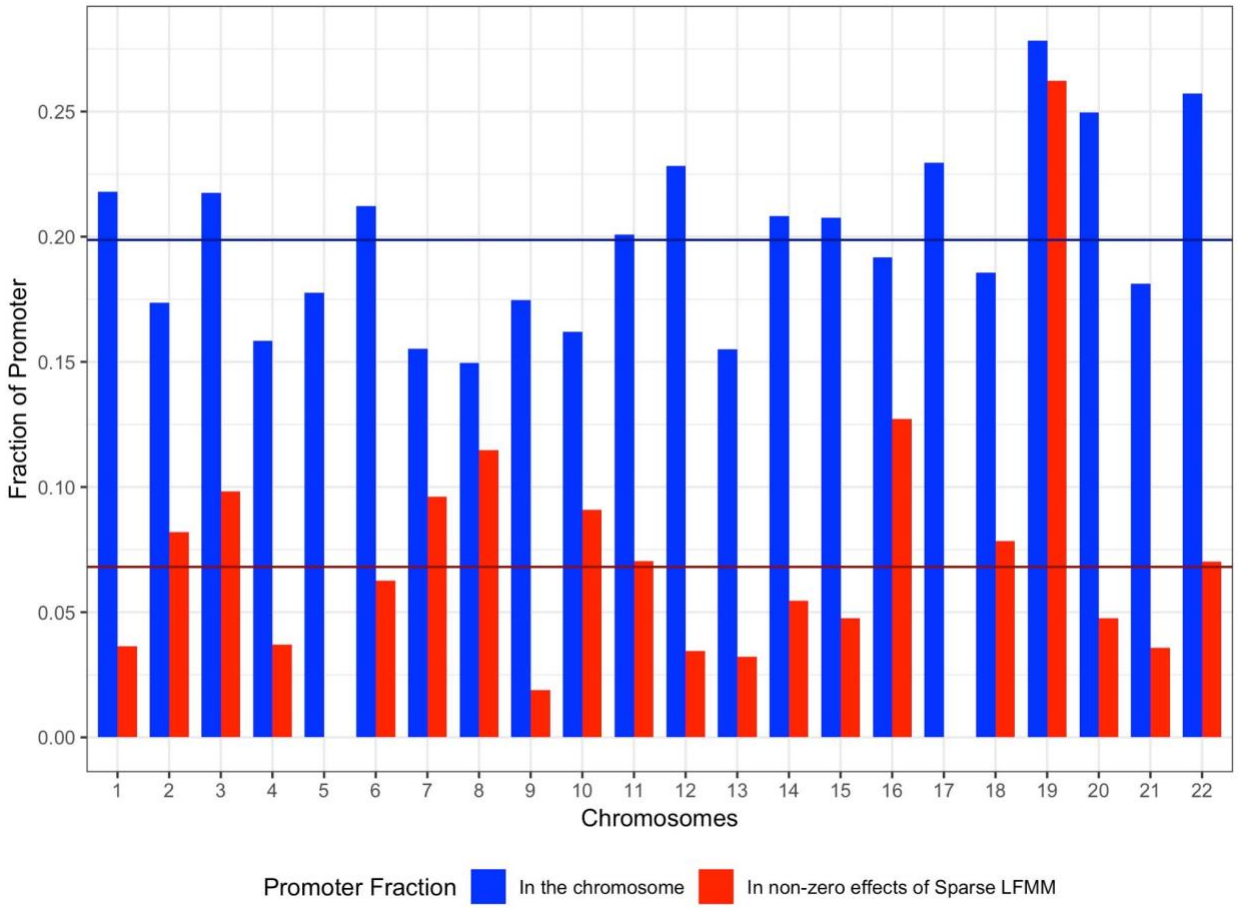

Figure 2.11 : EWAS sur le statut tabagique chez les femmes enceintes. *Sous-représentation des régions promotrices dans les hits de sparse LFMM par rapport à l'ensemble du méthylome. Les barres bleues correspondent à la proportion de régions promotrices dans chaque chromosome. Les barres rouges correspondent à la proportion de régions promotrices détectées par sparse LFMM. La ligne bleue horizontale représente la proportion moyenne de régions promotrices pour le méthylome. La ligne rouge représente la proportion moyenne pour sparse LFMM.*

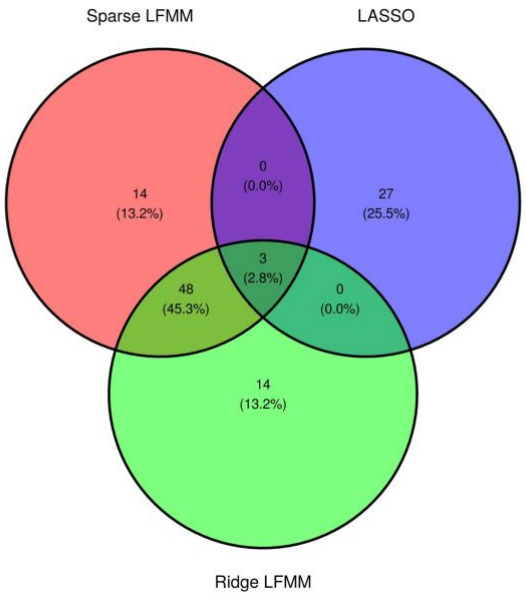

Figure 2.12 : EWAS sur le statut tabagique chez les femmes enceintes par Sparse LFMM, LASSO et ridge LFMM. *Diagramme de Venn des hits associés à la consommation de tabac par chaque approche. Pour sparse LFMM, les hits correspondent aux 5 % des tailles d'effet non nulles les plus élevées (65 hits). Pour LASSO, les hits correspondent aux 5% des tailles d'effet non nulles les plus élevées (30 hits). Pour ridge LFMM, les hits correspondent aux 65 CpGs ayant les tailles d'effet les plus élevées.*

2.4.2.3. Discussion

Nous avons appliqué sparse LFMM dans le cadre d'une EWAS, qui avait pour but d'identifier des marques de méthylation de l'ADN placentaire liées au tabagisme pendant la grossesse. La taille d'effet la plus forte, dans sparse LFMM mais aussi dans les "méthodes non-sparse", est située dans une région amplificatrice (cg27402634), proche du gène *LEKR1*. Ce gène a été associé, lors d'une précédente GWAS, au poids de naissance par le consortium Early Growth Genetics (EGG) (http://egg-consortium.org/birth-weight.html). Ce marqueur a également été détecté dans une étude indépendante portant elle aussi sur la méthylation placentaire et le tabagisme maternel

(Morales et al., 2016), et aussi dans une autre étude portant sur cohorte EDEN, qui contrairement à notre analyse, incluait les anciennes fumeuses dans leur population d'étude (Rousseaux et al., 2020). De plus, la liste des principaux résultats contenait plusieurs CpGs rapportés dans d'autres études similaires, notamment cg21992501 sur le gène *TTC27* (Cardenas et al., 2019), cg25585967 et cg17823829, respectivement dans les gènes TRIO et KDM5B (Everson et al., 2019; Morales et al., 2016).

Au sein de la liste complète des CpGs découverts, un enrichissement en régions amplificatrices et un appauvrissement en régions promotrices a été détecté, résultat en accord avec l'étude de (Rousseaux et al., 2020). Sparse LFMM nous a permis de confirmer les associations découvertes des précédentes études et de détecter de nouvelles associations incluant des gènes pour lesquels les changements de méthylation pourraient avoir des effets néfastes sur la santé de l'enfant.

2.5. Conclusion

Dans les études d'associations, il est extrêmement difficile de tenir compte de la variation due à des facteurs de confusion non observés. Nous avons donc introduit une nouvelle méthode, permettant de réaliser ce type d'étude, basée sur les facteurs latents. Sparse LFMM a comme nouveauté d'estimer à la fois les facteurs latents et les tailles d'effet des marqueurs. Ces tailles d'effet ont la particularité d'être "sparse" c'est à dire qu'une grande proportion d'effets sont nulles. L'estimation de facteurs latents permet de tenir compte de la possibilité qu'il y ait des facteurs de confusion non observés entre les marqueurs et l 'exposition (ou phénotype). Le caractère "sparse" des tailles d'effet vient de l'hypothèse qu'uniquement une faible proportion de tous les marqueurs soit corrélée avec l'exposition ou le phénotype.

Nous avons montré via des études de simulations et des analyses de données réelles, que notre méthode est performante, robuste et polyvalente car elle fonctionne aussi bien sur des données de génotype que sur des données de méthylation d'ADN. Étant donné le caractère "sparse" des tailles d'effet, notre méthode permet la sélection de variable sans avoir recours à des tests statistiques et donc sans avoir le problème des tests multiples.

L'application de notre méthode à des ensembles de données réelles a, d'une part, permis l'identification d'associations connues, et de l'autre, mis en évidence de nouvelles associations ayant des significations biologiques pertinentes.

3. HDMAX2: A framework for High Dimensional Mediation Analysis with application to maternal smoking, DNA methylation and birth outcomes

Ce chapitre reprend la trame de l'article « HDMAX2 : A framework for High Dimensional Mediation Analysis with application to maternal smoking, DNA methylation and birth outcomes » écrit par Jumentier Basile, Barrot Claire-Cécile, Maxime Estavoyer, Heude Barbara, François Olivier et Lepeule Johanna. L'article est en court de soumission.

Résumé :

L'analyse de médiation à haute dimension est une extension de l'analyse de médiation unidimensionnelle qui inclut plusieurs médiateurs et est de plus en plus utilisée en épidémiologie environnementale pour évaluer les effets épigénétiques indirects des expositions environnementales sur les résultats de santé. Cependant, les analyses impliquant des médiateurs de grande dimension soulèvent plusieurs problèmes statistiques. Bien que de nombreuses méthodes aient été récemment développées pour résoudre ces problèmes, aucun consensus n'a été atteint sur une combinaison optimale d'approches.

Nous avons développé HDMAX2, une nouvelle approche en plusieurs étapes de médiation qui combine des modèles de régression des facteurs latents pour les études d'association à l'échelle de l'épigénome avec des tests de médiation (test du maximum au carré). HDMAX2 a été soigneusement évaluée à partir de simulations et comparée à des méthodes de médiation en haute dimension. Ensuite, HDMAX2 a été utilisé pour évaluer les effets indirects de l'exposition au tabagisme maternel sur le poids à la

naissance à terme et l'âge gestationnel à l'accouchement dans une étude portant sur 470 femmes de la cohorte mère-enfant EDEN.

Durant les simulations HDMAX2 c'est montré plus puissante par rapport aux méthodes de médiation en haute dimension existantes. Elle a permis de détecter des régions non identifiées dans les analyses de médiation précédentes de l'exposition au tabagisme sur le poids de naissance. Les résultats ont fourni des preuves d'une architecture polygénique de la voie causale avec un effet indirect global de 44 g de poids corporel inférieur (31 % de la taille de l'effet total). HDMAX2 a également identifié des régions ayant des effets simultanés à la fois sur l'âge gestationnel et sur le poids de naissance. Parmi les principaux résultats des analyses de l'âge gestationnel et du poids de naissance, les régions situées sur les gènes *COASY* et *BLCAP* ont également médié la relation entre l'âge gestationnel et le poids de naissance, suggérant une causalité inverse dans la relation entre l'âge gestationnel et le méthylome.

Cette étude a mis en évidence plusieurs améliorations statistiques des analyses de médiation de haute dimension, qui ont révélé une complexité insoupçonnée des relations causales entre le tabagisme, le poids à la naissance et l'épigénome.

3.1. Introduction

Avec le développement des technologies de séquençage à haut débit, la modélisation causale en épidémiologie environnementale a pris une place importante pour évaluer les effets indirects de la méthylation de l'ADN (ADNm) dans la relation entre l'exposition et les effets sur la santé (Blum et al., 2020; MacKinnon et al., 2007; VanderWeele, 2015, 2016). Les effets du tabagisme maternel sur les issues de grossesse, telles que le faible poids à la naissance ou la prématurité, sont liés au développement ultérieur de maladies chroniques pendant l'enfance et à l'âge adulte.

Les preuves de plus en plus nombreuses de l'association entre le poids à la naissance et les maladies de l'adulte ont soutenu le concept des origines développementales de la santé future et des maladies (Barker, 2007; Gluckman et al., 2010). Les mécanismes causaux sous-jacents à ces associations entre expositions précoces et santé ultérieure sont cependant largement méconnus.

La méthylation de l'ADN, régulant l'expression des gènes sans modifier la séquence de l'ADN, est un mécanisme épigénétique largement étudié et est considéré comme un biomarqueur utile pour étudier les expositions environnementales prénatales ayant un impact sur la santé (Feinberg, 2018). Plus précisément, le tabagisme maternel pendant la grossesse a souvent été lié à des modifications de l'ADNm mesuré dans le sang de cordon (Agha et al., 2016; Küpers et al., 2015; Xu et al., 2021). L'ADNm placentaire a reçu moins d'attention que les autres tissus (Cardenas et al., 2019; Morales et al., 2016),

malgré son rôle clé dans la programmation fœtale. Le placenta soutient à la fois la santé de la mère et le développement du fœtus, notamment en fournissant des nutriments au fœtus et en régulant les échanges de gaz ou de déchets (Murphy et al., 2006). Comprendre les effets directs et indirects des modifications de l'ADNm dans les tissus placentaires sur le développement du fœtus et de l'enfant est un objectif pertinent.

L'évaluation des effets indirects de l'ADNm peut être réalisée grâce à une analyse de médiation, qui est un outil statistique pour comprendre les relations entre l'exposition et la santé grâce à l'inclusion d'une variable médiatrice située sur la voie causale entre les deux variables (Baron and Kenny, 1986). L'analyse de médiation de grande dimension est une extension de l'analyse de médiation unidimensionnelle, qui inclut de multiples médiateurs dans la voie entre l'exposition et la santé.

A l'heure actuelle, l'analyse de médiation en grande dimension classique pour des marqueurs ADNm repose sur une stratégie mixte utilisant plusieurs méthodes, et incluant trois étapes principales.

La première étape consiste à tester à la fois l'effet de l'exposition sur les niveaux d'ADNm et les effets des niveaux d'ADNm sur l'évènement de santé. La deuxième étape consiste à combiner les valeurs de significativité (*P*-valeur) obtenues à l'étape précédente afin de réduire le nombre de marqueurs candidats à la médiation. La dernière étape consiste à estimer leurs effets indirects sur l'évènement de santé.

Plusieurs approches d'analyse de médiation de grande dimension ont été proposées au cours des dernières années, et le développement de telles méthodes est toujours un domaine de recherche actif (Blum et al., 2020). Les développements récents entrent dans les catégories suivantes :

1) Réalisation d'analyses multidimensionnelles en réalisant des analyses de médiation unidimensionnelles pour chaque marqueur ADNm, par exemple en utilisant des tests de Sobel ou en estimant les Effets Causaux Moyens (ACME) (Imai et al., 2010; Sobel, 1982).

2) Amélioration des études d'association à l'échelle de l'épigénome (EWAS) réalisées avant l'analyse de médiation en incluant des estimations de la composition du type cellulaire en tant que covariables dans le modèle de régression (Teschendorff and Zheng, 2017). Dans ces analyses, la composition du type cellulaire est généralement estimée à partir de méthodes basées sur des références ou sans référence (Houseman et al., 2016, 2014; Rahmani et al., 2016).

3) Utilisation des régressions à facteurs latents dans les EWAS pour modéliser les effets de confusion non observés (Caye et al., 2019; Leek and Storey, 2007; Zhang et al., 2019) comme alternative à l'estimation de la composition des types cellulaires.

4) Amélioration des tests d'effets indirects de Sobel en combinant les valeurs de signification obtenues à partir des EWAS de diverses manières (Dai et al., 2020; Djordjilović et al., 2020, 2019; Sampson et al., 2018).

A ce jour, il n'existe pas de réel consensus sur une combinaison optimale des méthodes de chaque étape pour réaliser une analyse de médiation de grande dimension.

Dans cette étude, nous avons analysé le rôle causal des données d'ADNm placentaire dans l'association entre le tabagisme maternel pendant la grossesse et la croissance fœtale en utilisant une nouvelle méthode d'analyse de médiation de haute dimension (HDMAX2).

HDMAX2 évalue les effets et leurs valeurs de significativité (*P*-valeurs) entre l'exposition et l'ADNm et entre l'ADNm et l'évènement de santé. Pour réaliser ces deux associations nous utilisons des modèles de régression à facteurs latents. Ensuite, HDMAX2 combine les *P*-valeurs obtenues à partir des deux EWAS tout en contrôlant l'erreur de type I, afin de réaliser des tests de médiation. Nous avons également développé une extension à l'analyse de médiation avec HDMAX2 afin de détecter des régions différentiellement méthylées médiant la relation entre l'exposition et l'évènement.

A l'aide de simulations, nous avons effectué une évaluation approfondie des performances statistiques de HDMAX2 par rapport à un ensemble de méthodes utilisées dans les analyses de médiation en grande dimension sur des données ADNm. Enfin, nous avons utilisé HDMAX2 pour effectuer une analyse de médiation de grande dimension dans une étude portant sur $n = 470$ femmes de la cohorte mère-enfant EDEN (Heude et al., 2016). Nous avons identifié des marques de méthylation (CpG) placentaire et des régions génomiques médiatrices des effets du tabagisme maternel pendant la grossesse sur le poids et sur l'âge gestationnel à la naissance.

3.2. Analyse de médiation haute dimension (HDMAX2)

HDMAX2 (Figure 3.1) utilise deux régressions correspondant aux étapes standard d'une analyse de médiation unidimensionnelle (Baron and Kenny, 1986) étendue aux données ADNm multidimensionnelles. Nous avons développé une nouvelle approche en 3 étapes de l'analyse de médiation de haute dimension (HDMAX2) basée sur des modèles mixtes à facteurs latents, une méthode de combinaisons des *P*-valeurs de signification appelée test max-carré (max^2) et sur l'algorithme comb-p pour identifier des régions de CpGs (AMR pour Aggregate Methylated Region) médiant la relation entre l'exposition et l'évènement de santé.

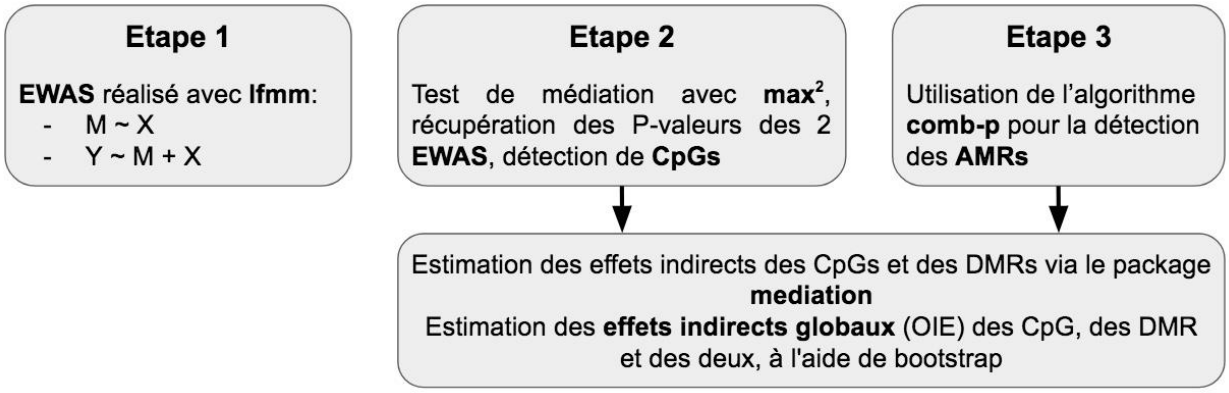

Figure 3.1 : Schéma synthétique de la méthode de médiation en haute dimension HDMAX2.

3.2.1. Étape 1 : Évaluer les associations entre l'exposition, les médiateurs et l'évènement de santé

HDMAX2 commence par une série de deux EWAS. Pour se faire nous utilisons les modèles mixtes à facteurs latents (LFMM) et notamment la version utilisant la pénalité « ridge » (Caye et al., 2019). Le premier EWAS consiste à tester les effets de l'exposition (*X*) sur la matrice (*M*) de marqueurs ADNm (CpG), ayant *n* lignes (échantillons) et *m* colonnes :

$$M = Xa_1^T + U_1V_1^T + E_1$$

Où a_1 contient les tailles d'effet de l'exposition sur les niveaux d'ADNm, U_1 est une matrice de *K* facteurs latents estimés simultanément avec a_1, V_1 contient les effets associés aux facteurs latents, et E_1 est une matrice d'erreurs résiduelles. Les facteurs latents *K* représentent des facteurs de confusion non observés, qui pourraient s'apparenter à des compositions de types cellulaires ou divers effets techniques liés à la mesure de la méthylation de l'ADN (effets de lot notamment) (Leek and Storey, 2008). L'inclusion de facteurs latents pour tenir compte de la composition des types cellulaires diffère des approches proposées précédemment basées uniquement sur des estimations des types cellulaires (Houseman et al., 2016; Rahmani et al., 2016; Teschendorff and Zheng, 2017), et ces facteurs latents sont donc plus généraux. Si elles sont disponibles, des covariables supplémentaires (c'est-à-dire l'âge maternel, le sexe du nouveau-né, ...) peuvent être incluses dans le modèle de régression. En utilisant la régression à facteur latent définie dans l'équation 1, une valeur de significativité, *P*x, est calculée pour le test d'effet de l'exposition sur l'ADNm ($H0 : a_{1,j} = 0$) pour le jième marqueur.

Un autre EWAS est effectué afin d'estimer les tailles d'effet pour l'association avec l'évènement de santé (*Y*) des niveaux d'ADNm comme suit :

$$M = Xa_2^T + Yb^T + U_2V_2^T + E_2$$

Où *b* contient les tailles d'effet des niveaux d'ADNm sur l'évènement de santé, U_2 sont des facteurs latents du modèle de régression à facteurs latents, V_2 sont les effets correspondants et E_2 est une matrice d'erreurs. Pour chaque marqueur j, une valeur de signification, *P*y, est calculée pour le test d'effet de l'ADNm sur l'évènement de santé (*H0 :* $b_j = 0$). A noter que les scores des tests des *P*-valeurs *Px* et *Py* sont recalibrés par le GIF (Genomic Inflation Factor) (Efron, 2004).

3.2.2. Étape 2 : Identifier les CpGs potentiellement médiateurs

La deuxième étape de HDMAX2 consiste à combiner les valeurs de significativité *P*x et *P*y calculées pour chaque marqueur ADNm en utilisant une nouvelle procédure appelée test du max-carré (ou max^2). La *P*-valeur pour le test du max-carré a été calculée de cette manière : $P = max(Px, Py)^2$. Comme le test de Sobel, le test du max-carré évalue l'hypothèse nulle selon laquelle l'effet de l'exposition sur l'ADNm ou l'effet de l'ADNm sur l'évènement de santé est nul. Le carré dans la formule garantit que la distribution des *P*-valeurs est uniforme lorsque *P*x et *P*y sont indépendants et uniformément distribués. Le test du max-carré a été utilisé pour réduire le nombre de marqueurs à évaluer en tant que médiateurs.

3.2.3. Étape 3 : Identifier les AMRs potentiellement médiatrices

Ici nous introduisons un nouveau concept : les AMRs (Aggregate Methylated Region). Nous définissons les AMRs comme un regroupement de CpGs médiant la relation entre une exposition et un évènement de santé. Pour détecter les AMRs nous avons adapté une méthode de détection de régions différentiellement méthylées (DMRs). Pour identifier les AMRs, nous avons utilisé comb-p, une méthode reposant sur la correction de Stouffer-Liptak-Kechris qui combine les *P*-valeurs des CpG adjacents au sein de fenêtres glissantes (Xu et al., 2016). Nous avons conservé les AMRs incluant au moins 2 CpGs, à une distance maximale de 1000 Pb et respectant un niveau de correction des test multiples (False Discovery Rate FDR) de 10 %. Pour résumer l'information des CpG situés dans les AMRs, une analyse en composantes principales a été réalisée sur les niveaux d'ADNm de ces marqueurs. Leur première composante principale a été retenue comme médiateur potentiel.

3.2.4. Quantification des effets indirects

Pour les CpGs sélectionnés à la 2e étape et pour les AMRs identifiées à la 3e étape, les estimations de l'effet indirect et les proportions médiées ont été calculées sur des données EDEN réelles en utilisant le package R "mediation" (Imai et al., 2010). Ce package a été utilisé sur chaque médiateur potentiel. A noter qu'au sein de chaque modèle de régression, nécessaire au package "mediation", nous incluons les covariables connues mais aussi les facteurs latents estimés par LFMM à l'étape 1. Les estimations

des effets indirects et des proportions médiées et leurs intervalles de confiances respectifs ont été réalisées à l'aide de 10000 simulations de MCMC.

Une nouveauté de HDMAX2 est d'évaluer un effet indirect cumulé pour tous les CpGs ou AMRs identifiés à l'étape 2 et 3. L'effet indirect global (OIE) a été estimé dans un modèle incluant m variables médiatrices comme suit :

$$y_i = c\, x_i + \sum_{j=1}^{m} b_j\, m_{ij} + \sum_{k=1}^{K} v_k\, u_{ik}^{(2)} + \varepsilon_i$$

Où (m_{ij}) représente les niveaux de méthylation observés à m médiateurs CpGs, AMRs, ou les deux, et les termes ($u_{ik}^{(2)}$) correspondent aux facteurs latents estimés à l'étape 1. L'effet indirect global a ensuite été calculé comme suit

$$OIE = \sum_{j=1}^{m} a_j\, b_j$$

Où (a_j) représente l'effet de l'exposition sur la méthylation (étape 1). Pour tenir compte de la corrélation entre les médiateurs, l'écart type de l'OIE a été calculé en utilisant une approche bootstrap (10 000 répétitions).

3.3. Utilisation de HDMAX2 sur des données simulées

Dans le but d'évaluer les performances de HDMAX2 et de ces différentes étapes nous avons conduit un ensemble de campagnes de simulations.

3.3.1. Méthodes

3.3.1.1. Évaluation de l'étape 1 : Évaluer les associations entre l'exposition, les médiateurs et les résultats via des modèles à facteurs latents

Au sein de HDMAX2 nous avons choisi d'utiliser la méthode LFMM pour estimer les facteurs latents mais ces facteurs latents pourraient être estimés par d'autres méthodes. Sur la base de notre ensemble de données EDEN (décrit en 3.4.1.1), une étude de simulation empirique a été réalisée pour décider laquelle des méthodes suivantes : "surrogate variable analysis" (SVA, (Leek and Storey, 2007)), "latent factor mixed models" (LFMM, (Caye et al., 2019)) et "confounder adjustment in multiple hypothesis testing" (CATE, (Wang et al., 2017)) fournit le meilleur algorithme d'estimation pour les facteurs latents dans notre analyse.

Pour conduire cette campagne de simulation, nous avons développé un simulateur conditionnel aux données de méthylation utilisées dans les études EWAS.

Considérons une matrice de profils de méthylation, M, avec n individus et m marqueurs. Nous supposons que les modèles de régression, $m_j = b_j X + E_j$, décrivent la relation entre une variable d'exposition non observée, X, et les niveaux de méthylation observés pour

le *j*-ième marqueur (b_j est la taille de l'effet de l'exposition sur la méthylation niveau *j*, et E_j a une variance σ_j^2). Conditionné à la matrice des profils de méthylation, on peut simuler l'exposition inobservée *X,* de variance σ_X^2 , de la manière suivante:

$$X|M = (m_1, \ldots, m_M) \sim \mathcal{N}\left(\frac{\sigma_x^2}{1 + \sigma_x^2 \left\|b/\sigma\right\|_2^2} \sum_{j=1}^{M} \frac{b_j m_j}{\sigma_j^2}, \frac{\sigma_x^2}{1 + \sigma_x^2 \left\|b/\sigma\right\|_2^2}\right)$$

Où

$$\left\|b/\sigma\right\|_2^2 = b_1^2/\sigma_1^2 + b_2^2/\sigma_2^2 + \ldots + b_M^2/\sigma_M^2.$$

En utilisant les données du chromosome 1 des données de méthylation placentaire de la cohorte EDEN, nous avons défini un ensemble de marqueurs causaux, pour lesquels b_j est non nul. Cent EWAS ont été générés à partir de ces données avec 30 marqueurs causaux choisis de manière aléatoire et σ_x = 0.18. Les tailles d'effet des marqueurs causaux ont été définies de la façon suivante : $b_j = \sigma_j / \sqrt{0.1}$. Le nombre de facteurs latents à estimer a été fixé à k = 5 pour SVA, CATE et LFMM. Les performances statistiques ont été évaluées par la précision (un moins le taux de fausses découvertes) et le F1-score (moyenne harmonique de précision et puissance).

3.3.1.2. Évaluation de l’étape 1 : Évaluer les associations entre l'exposition, les médiateurs et les évènements de santé via des modèles reposant sur l’estimation de la composition cellulaire

Pour évaluer l’apport des facteurs latents face aux estimations de la composition des types cellulaires, LFMM (inclus dans HDMAX2) a été comparée à deux modèles de régression RefFreeEWAS (Houseman et al., 2016) et Refactor (Rahmani et al., 2016), estimant des compositions cellulaires.

Pour répondre à cette problématique nous avons développé un simulateur de données de médiation en haute dimension. L'exposition et l’évènement (*X* et *Y*) et trois facteurs de confusion (*U*) ont été simulés selon un modèle gaussien multivarié. Le pourcentage de variance de l’exposition et de l’évènement expliqué par les facteurs latents a été fixé à 10%. Tout comme la corrélation entre *X* et *Y*. Les variances des facteurs latents sont égales à 1. Le nombre de marqueurs ADNm est égal à $m = 38\,000$, qui correspond approximativement au nombre de CpGs inclus dans le chromosome 1 dans nos données empiriques, et le nombre d'individus était égal à $n = 500$. Les vecteurs des tailles d'effet (*a* pour l'exposition et *b* pour le résultat) ont été générés en fixant une proportion de tailles d'effet à zéro. Des tailles d'effet non nulles ont été échantillonnées selon une distribution gaussienne standard. Une matrice d'erreur résiduelle *E* a été simulée en utilisant une distribution gaussienne multivariée avec des moyennes égales à 0 et des écarts types égaux à 1. En plus des trois facteurs latents, six facteurs supplémentaires représentant des types cellulaires ont été ajoutés dans le modèle de simulation. La proportion de types cellulaires a été simulée en utilisant une distribution de Dirichlet. Pour considérer des valeurs réalistes par rapport à notre analyse des données, les paramètres de la

distribution de Dirichlet ont été définis en utilisant les proportions de chaque type cellulaire estimées sur les données de méthylation de l'ADN placentaire EDEN (décrites plus loin). Une matrice de marqueurs ADNm a été construite en utilisant le modèle génératif ci-dessous :

$$M = Xa^T + Yb^T + UV^T + CT^T + E$$

Nous faisons varier 3 paramètres dans les simulations : la moyenne des tailles d'effet non nulle pour l'exposition (X) sur la méthylation M (a = 0.2, 0.4), la moyenne des tailles d'effet non nulle pour M sur l'évènement de santé (b = 0.2, 0.4) et le nombre de marqueurs causaux (égal à 8, 16 ou 32). Pour chaque ensemble de paramètres, 200 simulations ont été réalisées, et les méthodes de médiation ont été évaluées sur les données simulées. Pour chaque méthode testée, un sous-ensemble de marqueurs respectant un seuil FDR de 5 % (Benjamini and Hochberg, 1995) a été sélectionné comme médiateurs potentiels. Pour chaque liste de résultats, nous avons calculé la précision (1 - FDR), la sensibilité (puissance) et la moyenne harmonique de la précision et de la sensibilité (F1-score). La valeur la plus élevée d'un F1-score est un, si la précision et la sensibilité sont maximales, et la valeur la plus basse est zéro, si la précision ou la sensibilité est nulle.

En plus de la campagne de simulation visant à comparer LFMM avec RefFreeEWAS et Refactor. Nous avons réalisé une étude de simulation complémentaire permettant de comparer LFMM à une combinaison de LFMM avec RefFreeEWAS et Refactor. Pour ce faire nous avons préalablement estimé la composition cellulaire avec RefFreeEWAS ou Refactor et ensuite nous avons utilisé LFMM. Cette étude a été réalisée pour savoir s'il est possible d'optimiser les performances des méthodes en les combinant.

3.3.1.3. Évaluation de l’étape 2 : identifier les CpGs potentiellement médiateurs

Pour évaluer l’étape 2 de HDMAX2 qui consiste à sélectionner un sous-ensemble de médiateurs potentiels en utilisant le test du max-carré nous avons mené une nouvelle campagne de simulation utilisant le simulateur (et les simulations) décrit dans la partie "Évaluation de l’étape 1 : modèles reposant sur la composition cellulaire”.

Dans un premier temps, nous avons comparé HDMAX2 avec des méthodes basées sur l'application directe du test de Sobel ou l'analyse de médiation univariée (Cardenas et al., 2019; Morales et al., 2016). Ensuite, HDMAX2 a été comparée à un ensemble de méthodes récentes permettant des analyses de médiation en haute dimension sur des données de méthylation : HDMT est une procédure de tests multiples pour les hypothèses de médiation de haute dimension (Dai et al., 2020), ScreenMin est une procédure de contrôle du taux de fausses découvertes en deux étapes (Djordjilović et al., 2020), SBMH est une approche utilisant le contrôle du taux d'erreur par famille (FWER) et du taux de fausses découvertes (FDR) lors du test de plusieurs médiateurs (Sampson et al., 2018), un modèle de régression linéaire combiné à une ANOVA (Tobi et al., 2018) et une approche utilisant la sélection de variables pour réduire le nombre de médiateurs (Zhang et al., 2016) (HIMA). A noter que les méthodes HDMT, ScreenMin et SBMH fonctionnent de la même manière que le test max-carré, c'est-à-dire qu’elles combinent les séries de *P*-valeurs *Px* et *Py*. Pour ne pas avantager le test max-carré par rapport à ces 3 méthodes, nous avons choisi d’utiliser les *P*-valeurs *Px* et *Py* de la première étape de HDMAX2 dans ces 3 méthodes.

3.3.1.4. Évaluation de l'estimation de l'effet indirect global (OIE)

Nous avons conduit une dernière campagne de simulation visant à montrer que notre manière d'estimer l'OIE (Overall Indirect Effect) est correcte. Pour se faire nous avons construit un simulateur de données, simulant une exposition (*X*), des médiateurs (*m*) et un résultat de santé (*Y*) :

- Simulation de *X* de loi *N (0, 1)*
- Simulation de la taille d'effet (*c*) de *X* sur *Y*
- Simulation des tailles d'effet (*a*) de *X* sur les médiateurs (*m*)
- Simulation des 10 médiateurs de loi *N (0, 1)* en utilisant *X* et *a* ; $m_i = X * a_i + E$
- Simulation des tailles d'effet des médiateurs sur *Y (b)*
- Simulation de *Y* en utilisant *X*, *m* et *b* ; $Y = X * c + m_i * b_i + \ldots m_{10} * b_{10} + E$

Ou E représente une erreur gaussienne de loi *N (0, 1)*. De plus nous avons simulé un facteur de confusion qui est corrélé avec *X* et *Y* (*r = 0.1*).

Nous avons testé 10 scénarios (tableau 3) en faisant varier les tailles d'effet *a* et *b* et pour chaque scénario nous avons réalisé 100 simulations. A noter que pour chaque simulation, nous avons appliqué un coefficient multiplicateur (compris entre 0.1 et 5) aux tailles d'effet *a* et b. Pour chaque simulation, nous comparons, l'effet indirect global théorique (theoritical OIE) calculé sur les tailles d'effet *a* et *b* des simulations et l'effet indirect global estimé (estimate OIE).

Tableau 3.1 : Récapitulatif des scénarios de simulations.

Scénario	Effet	m1	m2	m3	m4	m5	m6	m7	m8	m9	m10
1	a	1	1	1	1	1	1	1	1	1	1
	b	1	1	1	1	1	1	1	1	1	1
2	a	1	1	1	1	1	1	1	1	1	1
	b	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
3	a	1	1	1	1	1	1	1	1	1	1
	b	-1	-1	-1	-1	-1	-1	-1	1	1	1
4	a	1	1	1	1	1	1	1	1	1	1
	b	1	1	1	1	1	1	1	-1	-1	-1
5	a	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	b	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
6	a	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	b	-1	-1	-1	-1	-1	-1	-1	1	1	1
7	a	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
	b	1	1	1	1	1	1	1	-1	-1	-1
8	a	-1	-1	-1	-1	-1	-1	-1	1	1	1
	b	-1	-1	-1	-1	-1	-1	-1	1	1	1
9	a	-1	-1	-1	-1	-1	-1	-1	1	1	1
	b	1	1	1	1	1	1	1	-1	-1	-1
10	a	1	1	1	1	1	1	1	-1	-1	-1
	b	1	1	1	1	1	1	1	-1	-1	-1

3.3.2. Résultats

HDMAX2 a été comparée à un large panel de méthodes (Figure 3.2). Les premières séries de simulations comparent les performances de LFMM à d'autres méthodes de régression pour estimer l'association entre l'exposition, les médiateurs et le résultat de santé (1ère étape de HDMAX2). Les deuxièmes séries de simulation comparent les performances du test max-carré à d'autres méthodes de médiation en haute dimension (2ème étape de HDMAX2).

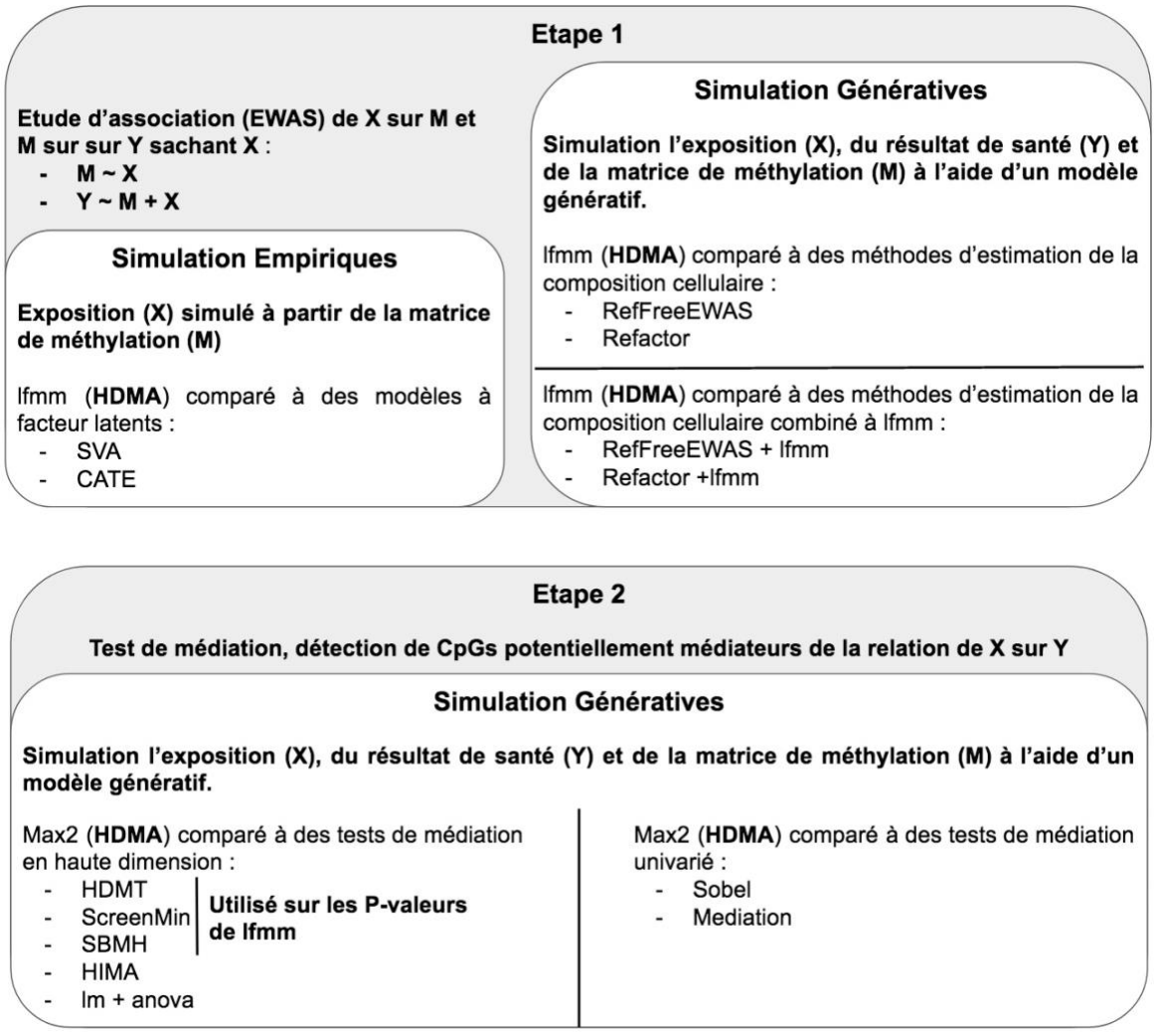

Figure 3.2 : Résumé des études de simulation réalisé pour évaluer les différentes étapes de HDMAX2.

3.3.2.1. Étape 1 de HDMAX2 : Évaluer les associations entre l'exposition, les médiateurs et les résultats

Ces campagnes de simulations ont pour but de tester la première étape de HDMAX2 qui consiste à utiliser LFMM pour réaliser 2 EWAS successives pour tester les associations entre l'exposition et l'ADNm et entre l'ADNm et l'évènement de santé.

3.3.2.1.1. Modèle à facteur latent

Sur la base de nos simulations empiriques d'EWAS, nous avons évalué les méthodes d'estimation des facteurs latents SVA, LFMM ou CATE. Que ce soit en termes de F1-score ou de précision, LFMM est la méthode qui obtient les meilleurs résultats (F1-score = 0.50 et précision = 0.68) (Figure 3.3). De plus, les temps d'exécution de LFMM sont plus courts que ceux de CATE.

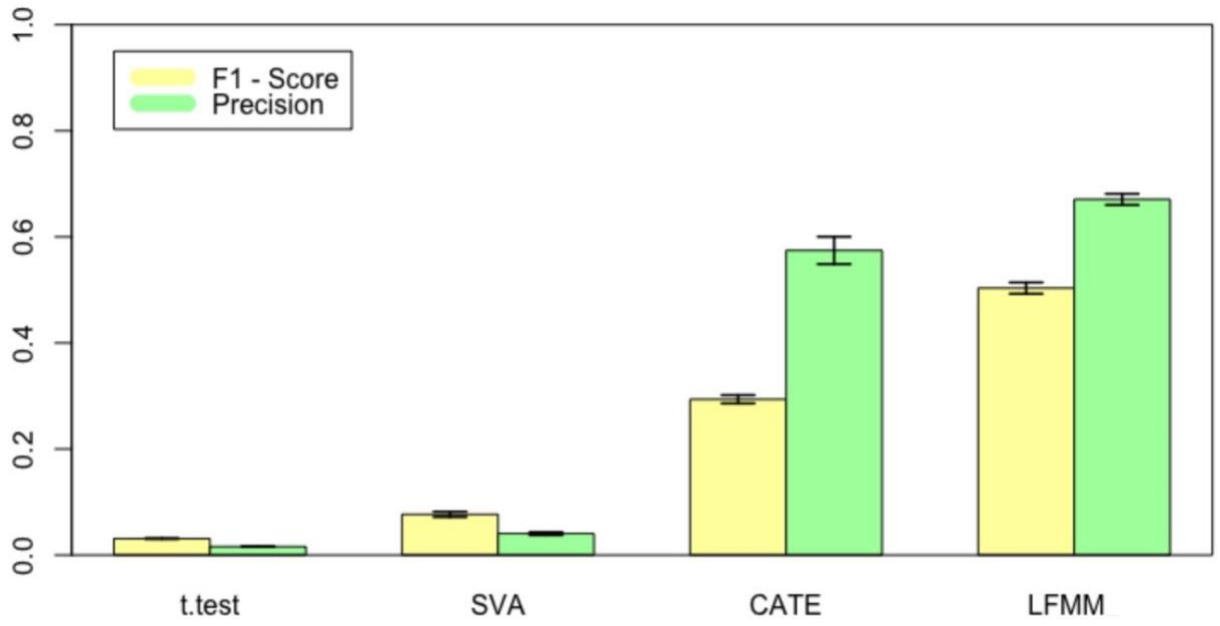

Figure 3.3 : Performances statistiques de quatre méthodes de régression pour réaliser des EWAS. *Les données d'exposition ont été simulées en fonction des données de méthylation d'ADN de la cohorte EDEN.*

3.3.2.1.2. Modèle de composition cellulaire

En utilisant notre simulateur de données de médiation en haute dimension, Nous avons comparé les performances relatives de LFMM à deux méthodes qui évaluent les compositions cellulaires, RefFreeEWAS et Refactor.

Dans tous les scénarios, les performances de Refactor étaient bien inférieures à celles de LFMM et RefFreeEWAS (Figure 3.4). Pour des faibles tailles d'effet de l'ADNm sur l’évènement de santé, les F1-scores étaient assez similaires entre LFMM et RefFreeEWAS, mais LFMM atteint des F1-scores plus élevés que RefFreeEWAS pour des tailles d'effet plus élevées. Toutes les approches ont obtenu des scores plus élevés lorsque le nombre de médiateurs était plus élevé, ou lorsque l'effet de l'exposition sur l'ADNm et l'effet de l'ADNm sur l’évènement étaient plus forts. Ces résultats indiquent que LFMM surpasse les méthodes qui tentent directement d'estimer la composition des types cellulaires à partir des données ADNm.

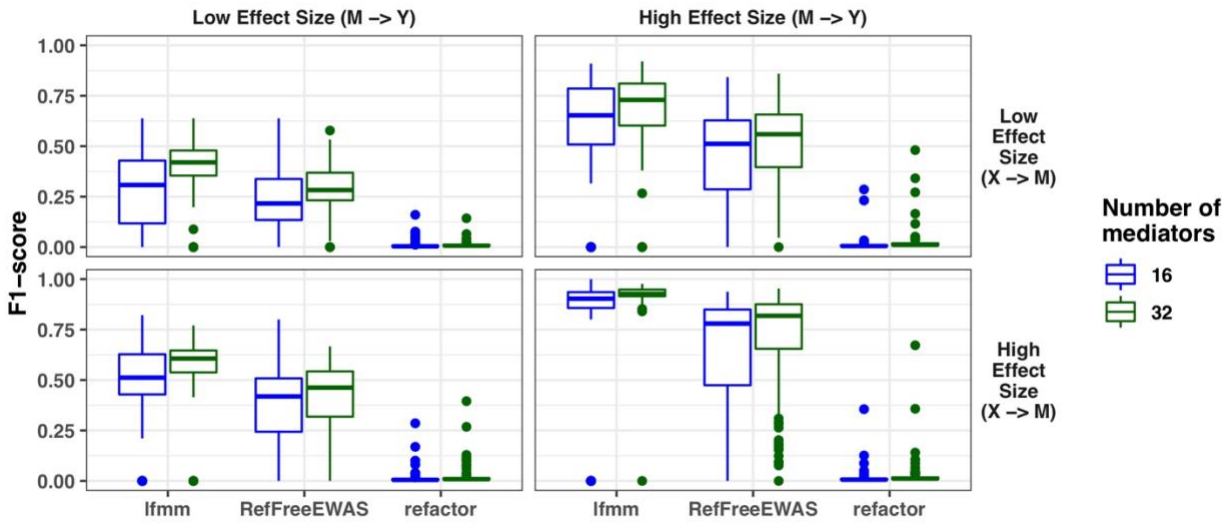

Figure 3.4 : Performances relatives des méthodes d'EWAS. *F1-score en fonction du nombre de médiateurs (16 ou 32), de la taille de l'effet de l'exposition sur l'ADNm (X -> M ; faible = 0.2, élevé = 0.4) et de la taille de l'effet de l'ADNm sur le résultat (M -> Y ; faible = 0.2, élevé = 0.4). Chaque simulation comprenait 38 000 CpG pour 500 échantillons, avec 6 types de cellules et 3 facteurs de confusion.*

Pour compléter ces résultats, nous avons réalisé une autre série de simulations combinant LFMM avec les méthodes d'estimations de la composition cellulaire (Figure 3.5). Comme pour la série de simulation précédente, l'ensemble des méthodes obtient de meilleures performances lorsque l'on a un plus grand nombre de médiateurs. Par contre, dans l'ensemble des scénarios, LFMM n'est que très légèrement supérieure aux combinaisons des méthodes : RefFreeEWAS + LFMM et Refactor + LFMM. Les moyennes générales des F1-scores sur l'ensemble des simulations sont de 0.5410 pour LFMM, 0.4890 pour RefFreeEWAS + LFMM et 0.4770 pour Refactor + LFMM. LFMM reste supérieure et nous conforte dans le choix de l'utiliser dans la procédure HDMAX2.

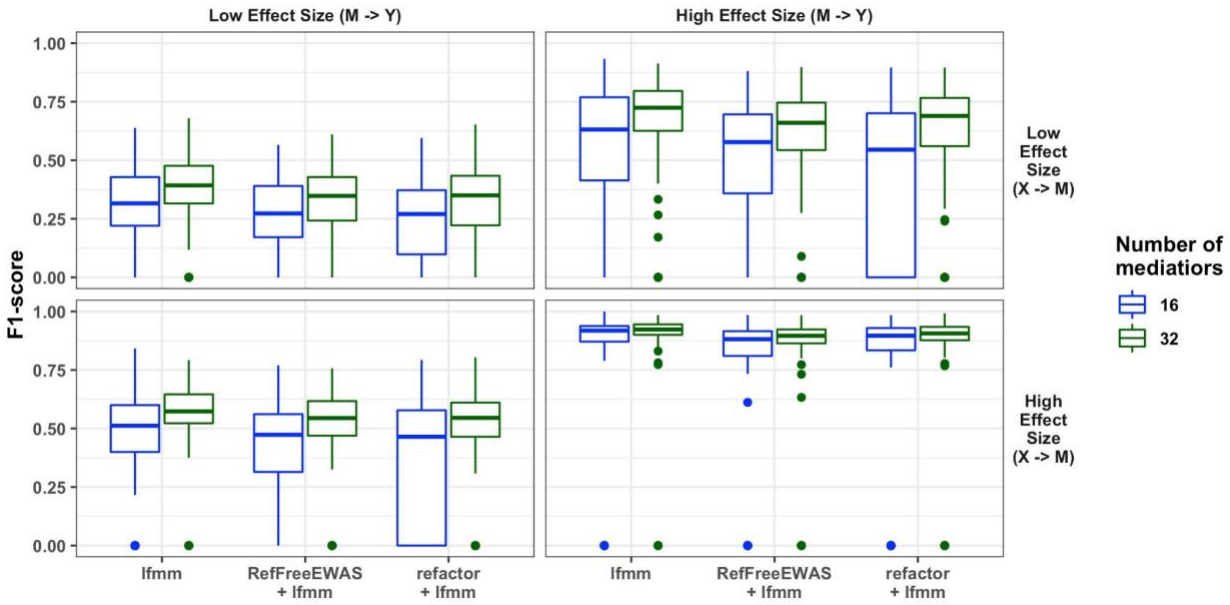

Figure 3.5 : Performances relatives des méthodes d'EWAS. *F1-score en fonction du nombre de médiateurs (16 ou 32), de la taille de l'effet de l'exposition sur l'ADNm (X -> M ; faible = 0.2, élevé = 0.4) et de la taille de l'effet de l'ADNm sur le résultat (M -> Y ; faible = 0.2, élevé = 0.4). Chaque simulation comprenait 38 000 CpG pour 500 échantillons, avec 6 types de cellules et 3 facteurs de confusion. Pour la méthode LFMM, nous utilisons k = 9. Pour la méthode RefFreeEWAS + LFMM, nous utilisons RefFreeEWAS k = 6 et LFMM k = 3. Pour la méthode Refactor + LFMM, nous utilisons Refactor k = 6 et LFMM k = 3.*

3.3.2.2. Étape 2 de HDMAX2 : identifier les CpGs potentiellement médiateurs

Dans un second temps, nous avons mené de nouvelles simulations pour évaluer les performances de notre seconde étape de HDMAX2 qui consiste à réaliser un test de médiation à l'aide du test de max-carré.

La première campagne de simulation avait pour but de comparer HDMAX2 à cinq méthodes récentes de médiation en haute dimension : HDMT, ScreenMin, SBMH, modèles linéaires combinés avec ANOVA (lm + anova) et HIMA (Figure 3.6). Dans chaque scénario, HDMAX2 et HDMT ont atteint des F1-scores quasi similaires, et ces approches étaient les meilleures dans l'ensemble. Dans le cas spécifique de fortes tailles d'effet de l'ADNm sur l'évènement de santé et de faibles tailles d'effet de l'exposition sur l'ADNm, lm + anova a obtenu les meilleurs scores, immédiatement suivis par HDMAX2 et HDMT. Les pires performances ont été obtenues par ScreenMin, SBMH et HIMA. Lorsque les deux tailles d'effet étaient élevées, HIMA a obtenu les performances les plus faibles. Pour des faibles tailles d'effet d'ADNm sur l'évènement de santé, lm + anova et SBMH ont obtenu les performances les plus faibles. De façon plus complète, nous avons aussi montré que HDMAX2 avait les temps de calculs les plus faibles notamment par rapport à HDMT (Figure 3.7).

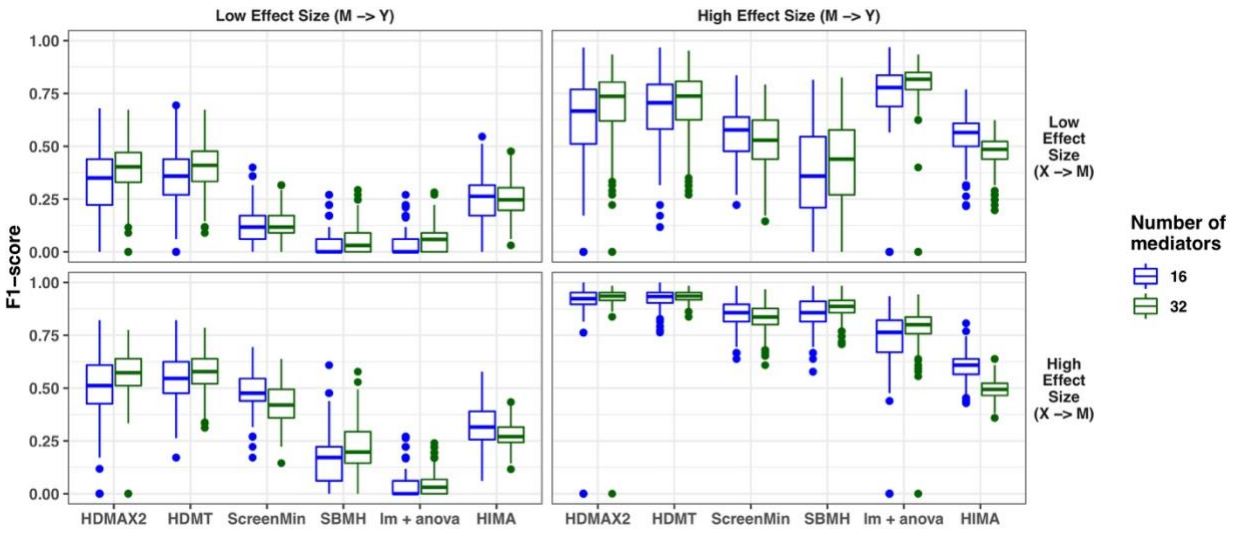

Figure 3.6 : Performances relatives des méthodes de médiation en haute dimension. *F1-score en fonction du nombre de médiateurs (16 ou 32), de la taille d'effet de l'exposition sur l'ADNm (X -> M ; faible = 0,2, élevé = 0,4) et de la taille d'effet de l'ADNm sur le résultat de santé (M -> Y ; faible = 0,2, élevé = 0,4). Chaque simulation comprenait 38 000 CpG pour 500 échantillons, avec 6 types de cellules et 3 facteurs de confusion supplémentaires.*

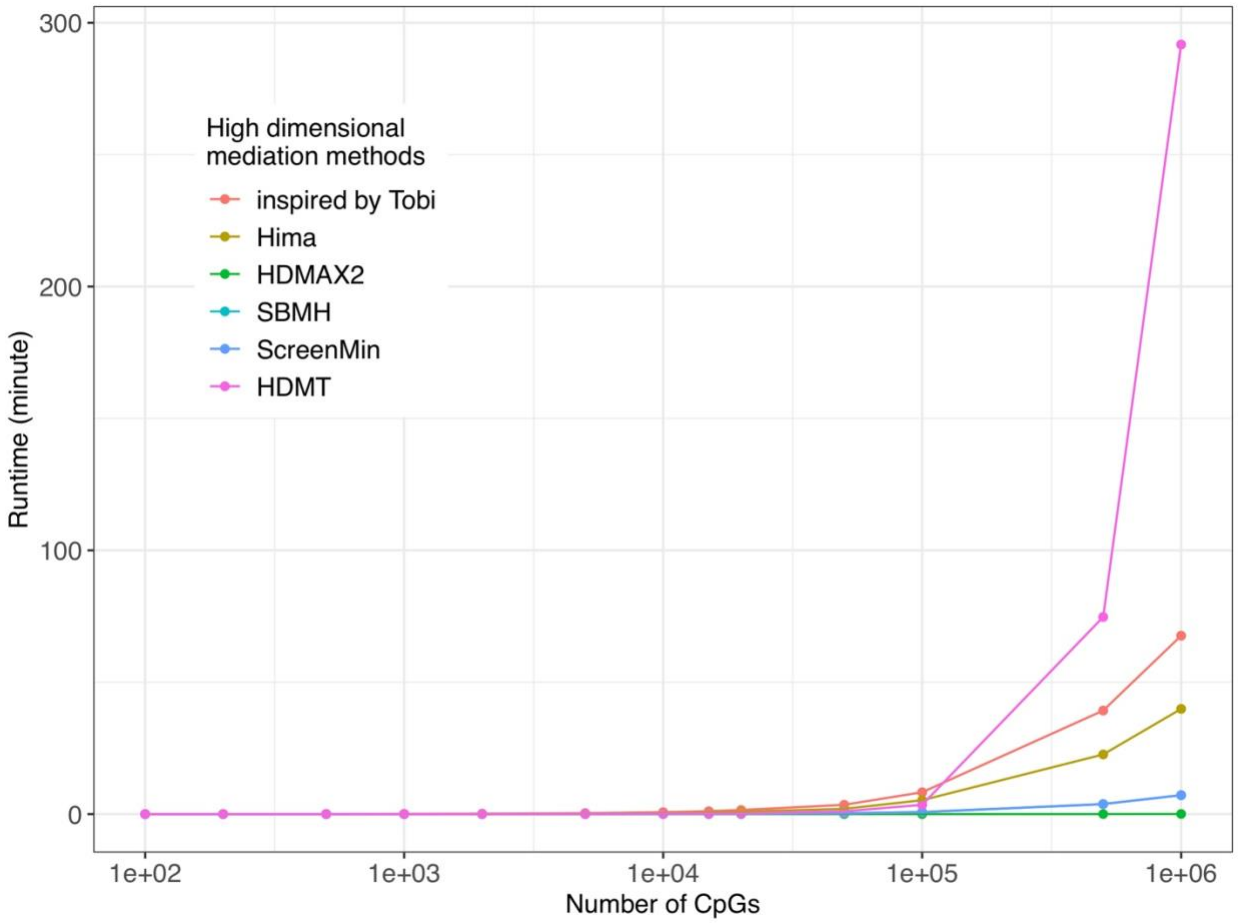

Figure 3.7 : Temps de calcul pour les méthodes de médiation de grande dimension en fonction du nombre de marqueurs (CpG). *Les temps d'exécution sont en minutes et le nombre de CpGs sont affichés en échelle logarithmique. Les courbes des méthodes ScreenMin et SBMH se superposent car leurs temps d'exécution sont égaux.*

Nous avons conduit une seconde étude de simulation visant à comparer le max-carré de HDMAX2 à des analyses de médiation combinant des EWAS avec des tests de Sobel et avec des analyses de médiation unidimensionnelles répétées pour chaque marqueur (Figure 3.8). Sur cet ensemble de simulation, HDMAX2 obtient de bien meilleures performances que les deux autres méthodes quel que soit le scénario.

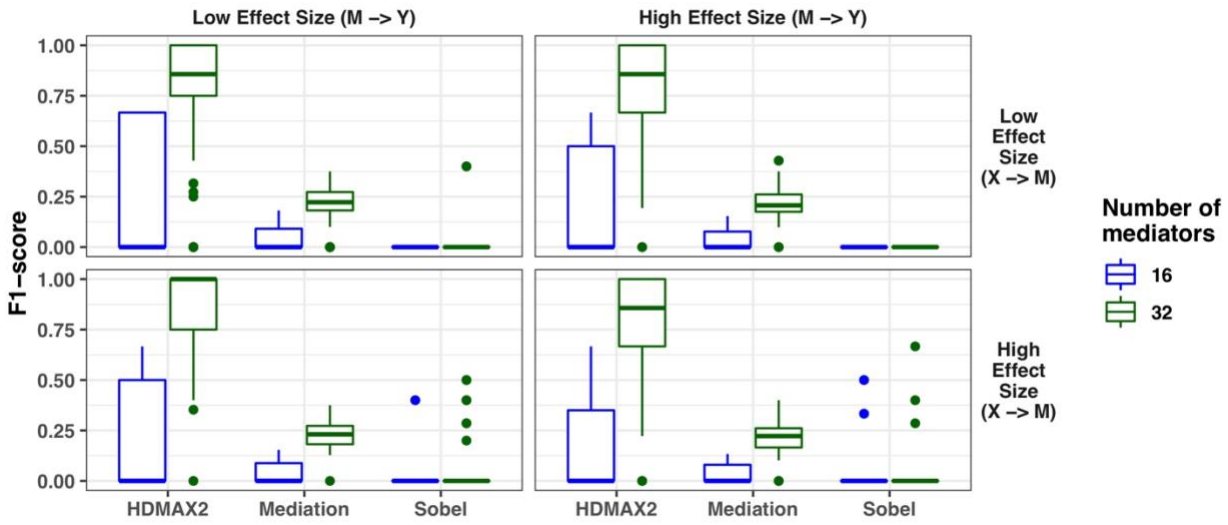

Figure 3.8 : Performances relatives de HDMAX2 face à des méthodes de médiations univariées. *F1-score en fonction du nombre de médiateurs (16 ou 32), de la taille d'effet de l'exposition sur l'ADNm (X -> M ; faible = 0,2, élevé = 0,4) et de la taille d'effet de l'ADNm sur le résultat de santé (M -> Y ; faible = 0,2, élevé = 0,4). Chaque simulation comprenait 38 000 CpG pour 500 échantillons, avec 6 types de cellules et 3 facteurs de confusion supplémentaires.*

3.3.2.3. Étape HDMAX2 d'estimation de l'effet indirect global

Nous avons voulu savoir si notre méthode de calcul de l'effet indirect global (OIE) par bootstrap était adapté (Figure 3.9). Nous avons donc conduit un ensemble de simulation ayant 10 scénarios différents. Au sein de ces 10 scénarios, nous avons fait varier la taille des effets de X sur les médiateurs et la tailles d'effets des médiateurs sur Y. Nous avons comparé les estimations d'OIE à leurs valeurs théoriques. Les résultats des simulations sont très similaires entre les scénarios. D'une part les estimations d'OIE sont très corrélées avec les valeurs théoriques ($r > 0.98$) et les estimations d'OIE sont environs 20% plus faibles que les valeurs théoriques.

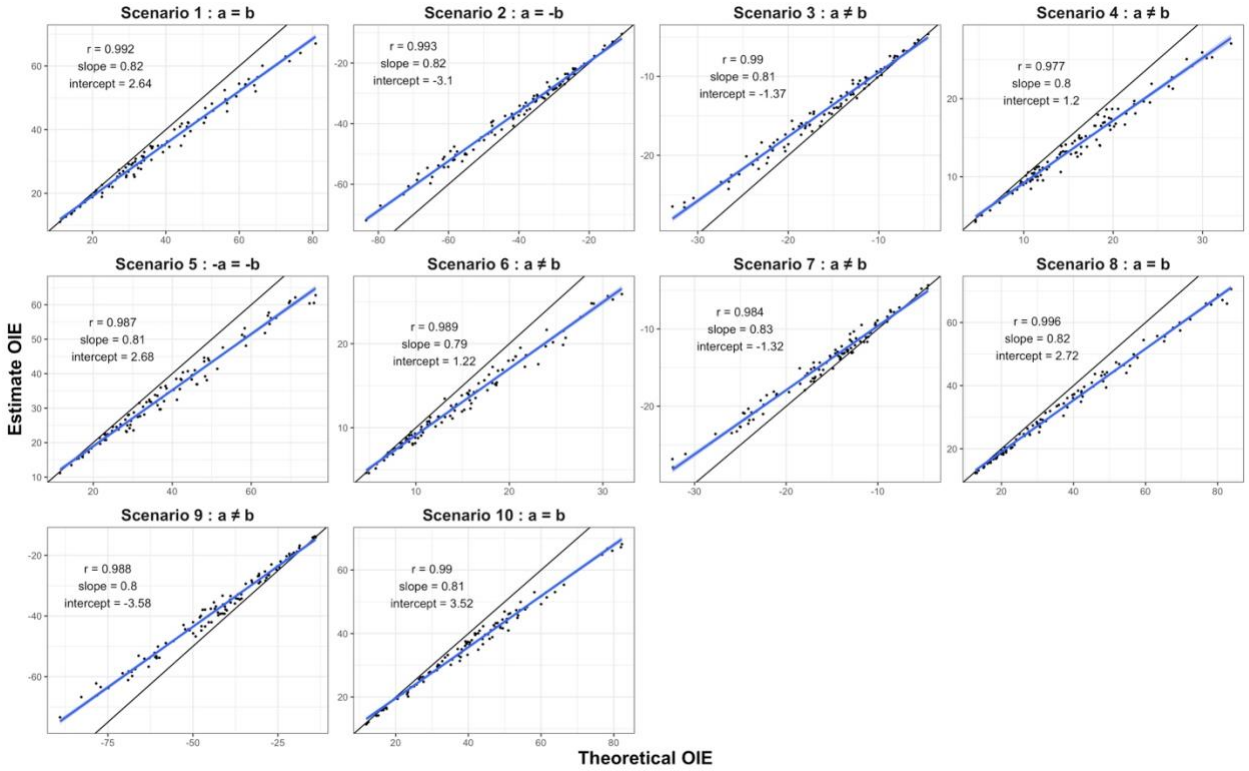

Figure 3.9 : Effet indirect global estimé (Estimate OIE) en fonction de l'effet indirect global théorique (Theoretical OIE) dans différent scénarios de simulation.

Pour conclure, nous avons pu montrer que nos estimations d'OIE sont robustes quel que soit le scénario de simulation, qu'elles sont fortement corrélées aux valeurs théoriques mais que nous sous-estimons légèrement leurs valeurs par rapport aux valeurs théoriques.

3.3.3. Résumé de HDMAX2

On a récemment découvert que l'ADNm placentaire jouait un rôle important dans la relation causale entre le tabagisme maternel pendant la grossesse et la croissance fœtale (Cardenas et al., 2019; Morales et al., 2016; Nakamura et al., 2021). Cette étude avait donc comme objectif de comprendre la relation complexe entre le tabagisme maternel durant la grossesse, la méthylation de l'ADN placentaire et le poids à la naissance. A la vue du manque de méthodes pour réaliser des analyses de médiation en haute dimension, nécessaires à notre étude, nous avons développé une nouvelle procédure appelée HDMAX2 (pour High Dimensional Mediation Analysis). Cette méthode à deux étapes, repose sur une combinaison intelligente d'une méthode EWAS (LFMM), utilisant des modèles à facteurs latents, et d'un test de médiation en haute dimension (max-carré). De plus, notre méthode permet la détection de régions pouvant médier la relation entre l'exposition et l'évènement de santé.

Dans le but de tester et d'évaluer notre méthode, nous avons conduit un ensemble de campagnes de simulations. Les premières campagnes de simulations avaient pour but de tester l'étape 1 de HDMAX2 qui utilise LFMM tandis que les suivantes avaient pour but de tester la seconde étape (max-carré).

Nous avons comparé LFMM à un ensemble de méthodes utilisant des facteurs latents pour réaliser des EWAS. Cette campagne de simulation était basée sur un simulateur de données reprenant les données de ADNm de la cohorte EDEN. Nous avons pu montrer que LFMM était la méthode la plus performante.

Ensuite, sur la base d'un nouveau simulateur de données, permettant de simuler des données de médiation en haute dimension, nous avons évalué les performances de LFMM face à un ensemble de méthodes basées sur l'estimation de la composition cellulaire pour réaliser des EWAS. Cette simulation a permis de mettre en évidence la supériorité de LFMM face à RefFreeEWAS et Refactor. De plus, même une combinaison de RefFreeEWAS et Refactor avec LFMM ne montrait pas de résultats supérieurs à LFMM.

Concernant l'évaluation de l'étape 2 de HDMAX2, en premier nous avons montré que HDMAX2 était supérieure à des méthodes de médiation unidimensionnelle appliquées à des données de grande dimension. Pour finir, nous avons conduit une dernière campagne de simulation comparant notre test de médiation max-carré à un ensemble de méthodes de médiation en haute dimension. Ces simulations ont mis en évidence deux méthodes : max-carré (HDMAX2) et HDMT, qui sont les méthodes ayant obtenu les meilleures performances. Ces deux méthodes combinent des *P*-valeurs pour réaliser leur test de médiation. Pour simplifier nos simulations, nous avons choisi d'utiliser HDMT avec les *P*-valeurs de l'étape 1 de HDMAX2 (comme pour max-carré). Étant donné que HDMAX2 et HDMT ont obtenu les meilleures performances, ce résultat nous conforte dans l'utilisation de LFMM à la première étape de HDMAX2. Pour finir nous avons montré que les temps de calcul de HDMAX2 étaient bien plus faibles que ceux de HDMT si on considère un très grand nombre de marqueurs à tester (> 500K).

3.4. Impact du tabagisme maternel sur le poids et l'âge gestationnel à la naissance via la méthylation de l'ADN placentaire

Dans cette partie, nous appliquons notre nouvelle procédure HDMAX2 pour comprendre les relations entre le tabagisme maternel, le méthylation de l'ADN placentaire et le poids à la naissance. Il est connu que le tabagisme maternel entraine une diminution du poids de naissance qui est lui-même lié à l'apparition de maladies cardiovasculaires à l'âge adulte (Rich-Edwards et al., 1997). Notre hypothèse est qu'une partie des effets néfastes du tabac sur le poids de naissance soient régulés par la méthylation de l'ADN.

3.4.1. Méthodes

3.4.1.1. Données de la cohorte mère-enfant EDEN

Notre analyse de médiation en haute dimension visant à comprendre l'impact du tabagisme sur le poids à la naissance via la méthylation placentaire s'appuie sur les données de la cohorte EDEN. Les participants de cette cohorte mère-enfant ont été inclus dans les hôpitaux universitaires de Nancy et Poitiers, France, entre 2003 et 2006 (Abraham et al., 2018; Heude et al., 2016). Des données sur le mode de vie, démographiques et médicales ont été recueillies par des questionnaires et des entretiens pendant la grossesse et après l'accouchement.

La puce "BeadChip Infinium HumanMethylation450" d'Illumina a été utilisée pour mesurer la méthylation à partir d'ADN extrait de 668 échantillons de placenta. Les protocoles d'extraction d'ADN placentaire et de traitement de l'ADNm sont détaillés dans (Jedynak et al., 2021). Les données d'ADNm ont été normalisées à l'aide de la méthode BMIQ

(pour Beta-Mixture Quantile) pour finalement obtenir des niveaux de bêta-méthylation pour 379 904 sites CpG (Teschendorff et al., 2013).

Sur les 668 femmes qui ont vu leur ADNm placentaire mesuré, nous avons exclu les accouchements prématurés (n=28), les anciennes fumeuses (n=70) et les femmes dont le statut tabagique était inconnu (n=100) et nous arrivons donc à 470 femmes.

Le poids à la naissance a été extrait des dossiers médicaux. Le tabagisme maternel prénatal a été recueilli par questionnaires lors des examens cliniques pré et post-nataux. *Les non-fumeuses* ont été définies comme les femmes n'ayant pas fumé pendant les 3 mois précédant et pendant la grossesse (359 non-fumeuses). *Les fumeuses* étaient définies comme des femmes fumant plus d'une cigarette par jour pendant toute la durée de la grossesse (111 fumeuses). Toutes les fumeuses pendant la grossesse ont également fumé pendant les 3 mois précédant la grossesse.

3.4.1.2. Application de l'analyse de médiation en haute dimension (HDMAX2)

Nous avons émis l'hypothèse que le tabagisme maternel pendant la grossesse pourrait induire des modifications de l'ADNm placentaire entraînant des modifications de l'âge gestationnel ou du poids à la naissance. À cette fin, nous avons étudié les relations causales entre le tabagisme maternel, l'ADNm placentaire et les 2 évènements de santé. Le tabagisme maternel a été codé en variable catégorielle et les évènements de santé ont été codés en variables continues. HDMAX2 a d'abord été utilisé pour détecter des CpGs potentiellement médiateurs puis pour détecter des régions différentiellement méthylées (AMR) pouvant médier la relation.

Dans les modèles de régression HDMAX2, nous avons inclus des facteurs d'ajustement pouvant être des facteurs de confusion: sexe de l'enfant, parité (0, 1, ≥ 2 enfants ; variable catégorielle), âge de la mère à la fin des études (≤ 18, 19-20, 21-22, 23-24, ≥ 25 ans ; variable catégorielle), saison de conception (variable catégorielle), centre d'inclusion, indice de masse corporelle (IMC) de la mère avant la grossesse, âge maternel à l'accouchement, facteurs techniques liés aux mesures de l'ADNm (lots et puces, variables catégorielles). Nous nous sommes appuyés sur l'analyse en composantes principales (ACP) de la matrice ADNm pour inclure 6 facteurs latents dans les modèles de régression HDMAX2 (Figure 3.10). Ce nombre était cohérent avec les 6 facteurs sélectionnés dans les travaux précédents sur la cohorte EDEN (Abraham et al., 2018; Rousseaux et al., 2020), mais dans ces travaux ils utilisaient la méthode RefFreeEWAS pour estimer des compositions cellulaires et non des facteurs latents.

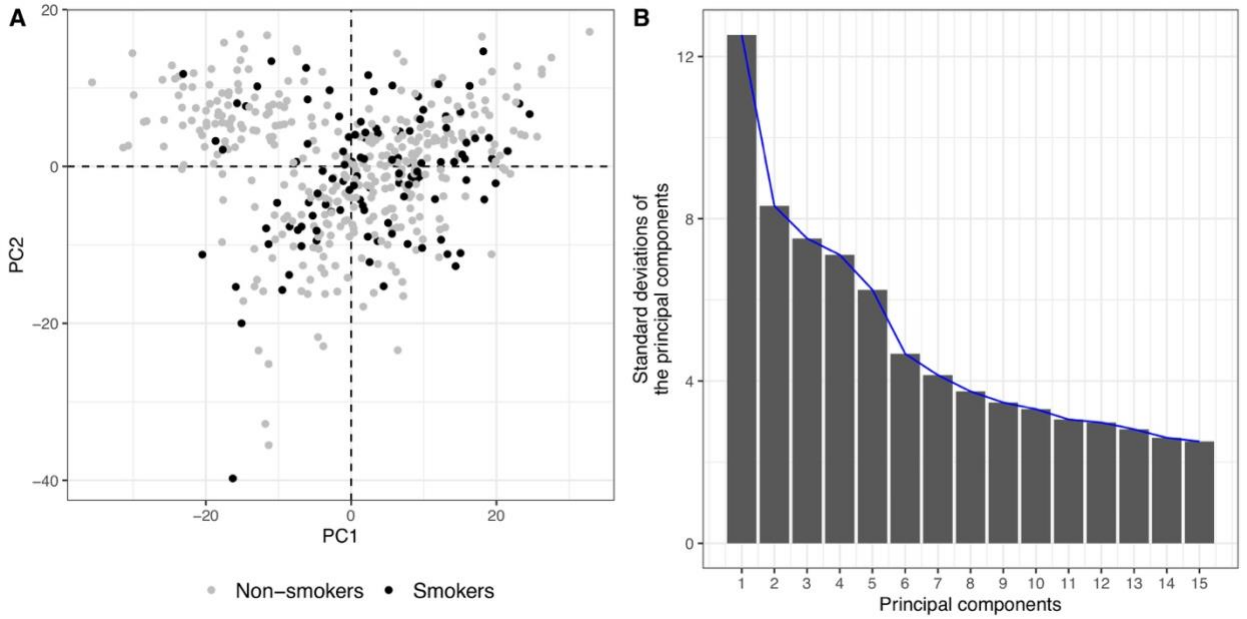

Figure 3.10 : Analyse en composante principale des données de méthylation d'ADN placentaire de la cohorte EDEN (n = 470).

Les *P*-valeurs de HDMAX2 ont été transformées en Q-valeurs par l'algorithme de FDR local présent dans le package R "fdrtool" (Strimmer, 2008) ce qui nous permet de contrôler le FDR.

La calibration des *P*-valeurs *des* 379 904 CpG du test a été évaluée par un examen direct de l'histogramme des *P*-valeurs (figure 3.11). De plus, le eta0 a été calculé pour évaluer la proportion d'hypothèse nulle parmi les 379 904 tests. Cette proportion a été estimée à eta0 = 99,8-99,9 % (BW-GA), ce qui suggère qu'un niveau de FDR de 5 % serait trop conservateur. Pour être en accord avec la valeur de eta0, les CpG candidats ont été sélectionnés à des niveaux de FDR < 10 %, correspondant à une *P*-valeur < $9{,}03 \times 10^{-6}$ pour le poids de naissance et à une *P*-valeur < $3{,}27 \times 10^{-6}$ pour l'âge gestationel. Les

résultats obtenus après avoir considéré les niveaux de FDR < 20 % sont également rapportés.

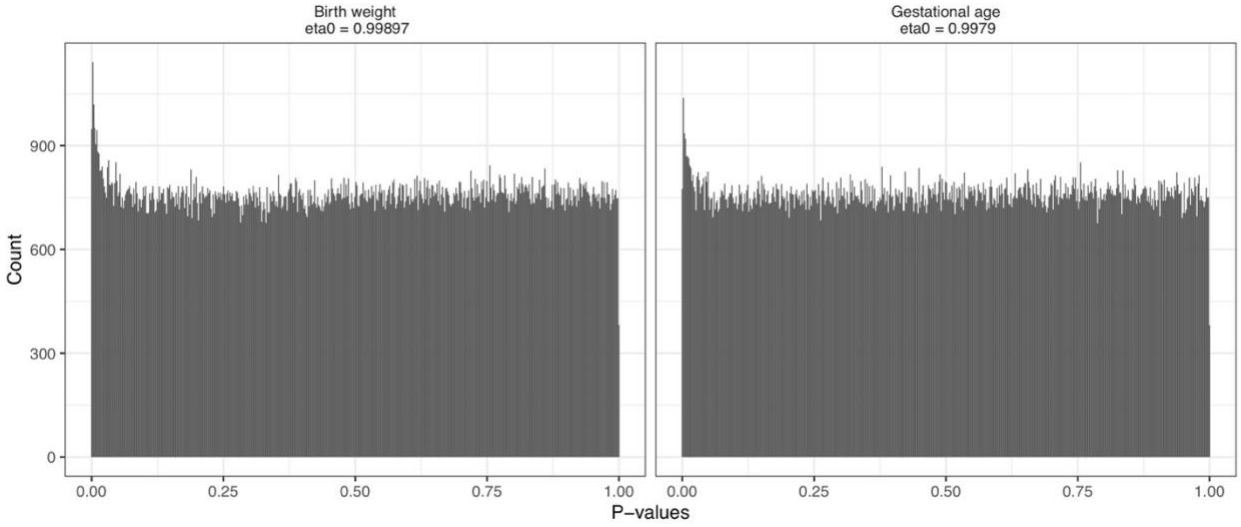

Figure 3.11 : Histogramme des P-valeurs de HDMAX2 pour les analyses de Poids à la naissance (Birth weight) et d'âge gestationnel (Gestational age).

Pour ce qui est des analyses bio-informatiques, les régions de promoteur et d'enhancer sur les CpG ont été obtenues à partir des annotations de la puce Illumina. Les annotations génétiques ont été obtenues à l'aide du package FDb.InfiniumMethylation.hg19 (Triche, 2014).

Nous avons comparé l'expression des gènes dans le placenta par rapport aux autres tissus en utilisant la base de données Expression Atlas (Papatheodorou et al., 2020). Pour chaque gène, le critère de Chauvenet a été utilisé pour décider si l'expression du gène dans le placenta était différente par rapport à d'autres tissus. L'annotation fonctionnelle a été faite à partir des bases de données KEGG et Gene Annotation (Kanehisa et al., 2021).

Pour finir, dans l'optique de valider notre procédure HDMAX2, nous avons également réalisé l'analyse de médiation « tabagisme maternel – méthylation – poids à la naissance » et « tabagisme maternel – méthylation – âge gestationnel » avec différentes combinaisons de méthodes. Pour les EWAS, nous avons choisi de les réaliser avec LFMM (K = 6) ou avec RefFreeEWAS (K = 6). Comme pour LFMM, les *P*-valeurs de RefFreeEWAS ont été calibrés en utilisant le GIF (Genomic Inflation Factor). Ensuite pour réaliser les tests de médiations, nous avons utilisé HDMT. RefFreeEWAS et HDMT ont été utilisés car ces deux méthodes montraient de bonnes performances dans les études de simulations. Au final, HDMAX2 a été comparé aux combinaisons de méthodes suivantes : LFMM + HDMT, RefFreeEWAS + max2 et RefFreeEWAS + HDMT. Pour chacune de ces méthodes nous avons regardé les rangs des CpGs en fonction de leurs *P*-valeurs.

3.4.2. Résultats

Nous avons cherché à comprendre l'impact du tabagisme sur le poids à la naissance via la méthylation, et notamment nous avons cherché si des CpGs ou des régions de CpGs pouvaient être impliqués. Pour ce faire, à l'aide de notre méthode HDMAX2, nous avons réalisé une analyse de médiation en haute dimension sur la cohorte EDEN. De plus, pour compléter l'analyse de l'impact du tabagisme maternel sur l'ADNm placentaire nous avons conduit une seconde analyse de médiation en haute dimension sur l'âge gestationnel à la naissance.

Parmi les 470 couples mère-enfant, l'âge moyen des mères était de 29 ans (sd = 5 ans), l'indice de masse corporelle moyen de la mère avant la grossesse était de 22,98 kg/m ^{2} (sd = 4,38 kg/m^2) et 23,6 % des femmes fumaient pendant la grossesse. (Tableau 3.1). Le poids de naissance à terme variait de 2010g à 4960g, avec une moyenne de 3352g (sd = 435). L'âge gestationnel variait de 37 à 42 semaines d'aménorrhée, avec une moyenne de 40 semaines (sd = 1,20 semaine). Le tabagisme maternel pendant la grossesse était corrélé significativement au poids à la naissance (r = -0,16, *P* = 0,003), mais pas à l'âge gestationnel (Figure 3.12). Le poids à la naissance et l'âge gestationnel étaient significativement corrélés (r = 0,31, *P* = 1,57 x 10 $^{-12}$). Après ajustement, l'effet total du tabagisme maternel était de -140 g sur le poids à la naissance (sd = 49,1 g, *P* = 0,004) et cet effet n'était pas significatif pour l'âge gestationnel (P = 0,24).

Tableau 3.1 : Caractéristiques de la cohorte mère-enfant EDEN (n = 470).

Categorical variable		Count	(%)	Continuous variable	mean	(sd)
Centre						
	Nancy	281	59.8	BMI (kg/m2)	22.98	4.38
	Poitiers	189	40.2			
Sex of offspring				Gestational Age (weeks)	39.99	1.2
	Male	241	51.3			
	Female	229	48.7	Maternal Age (year)	29.38	4.99
Parity						
	0	189	40.2	Birth Weight (g)	3351.9	435
	1	195	41.5			
	2	86	18.3			
Maternal age at end of education (year)						
	<18	89	18.9			
	19-20	70	14.9			
	21-22	114	24.3			
	23-24	109	23.2			
	>25	87	18.5			
Season of conception						
	January–March	100	21.3			
	April–June	103	21.9			
	July–September	130	27.7			
	October–December	137	29.1			
Maternal Smoking						
	Smoker	111	23.6			
	Non-smoker	359	76.4			

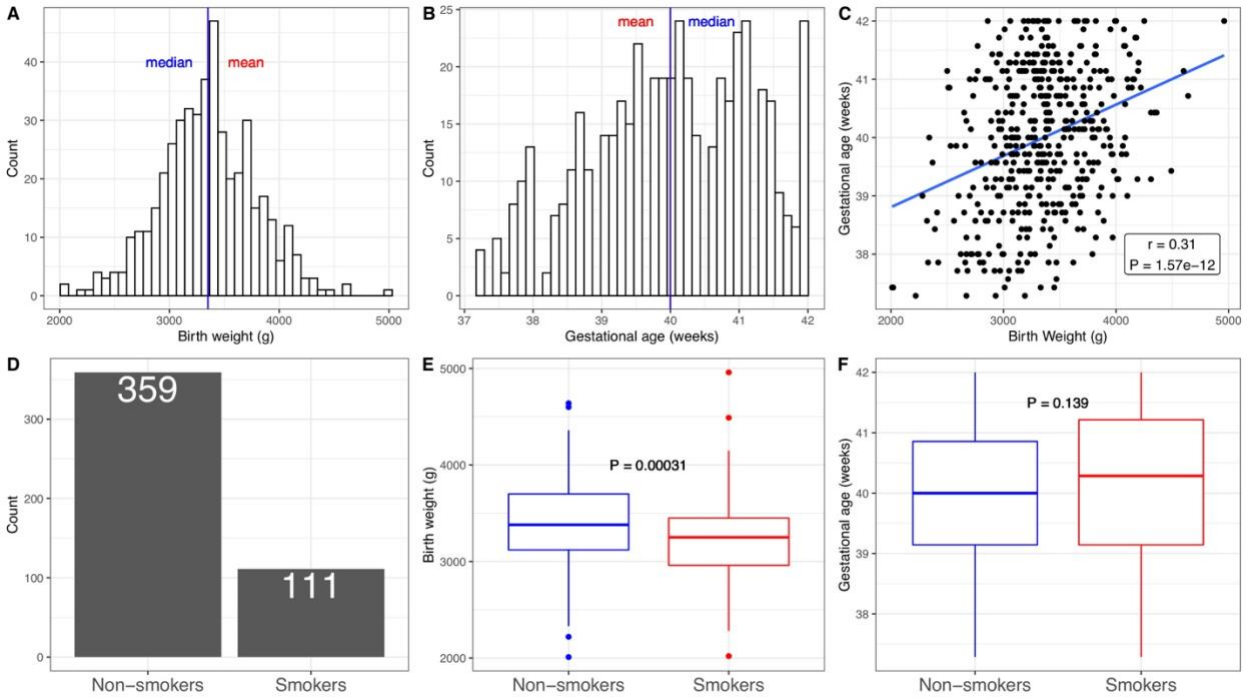

Figure 3.12 : Description des données d'exposition et des résultats de santé dans la cohorte mère-enfant EDEN. *A) Distribution des poids à la naissance. B) Distribution de l'âge gestationnel. C) Poids de naissance en fonction de l'âge gestationnel. D) Répartition des fumeurs et des non-fumeurs. E) Distribution des poids de naissance pour les fumeurs et les non-fumeurs. F) Distribution de l'âge gestationnel pour les fumeurs et les non-fumeurs.*

3.4.2.1. Analyse du poids de naissance

HDMAX2 a été utilisé pour l'analyse de médiation en haute dimension du tabagisme maternel sur le poids de naissance via la méthylation d'ADN placentaire. À un niveau FDR de 10 %, 32 CpGs ont été identifiés comme médiateurs (figure 3.13, *Pmax-carré* ajusté < 9,03 x 10^{-6}), (dix-neuf CpG à un niveau FDR de 5 %).

20 CpGs étaient associés à une diminution du poids à la naissance (ACME moyen : -32,1 g, SD = 5,4 g ; proportion médiée (PM) moyenne : 24 %, SD = 0,07) tandis que 12 CpGs étaient associés à une augmentation du poids de naissance (ACME moyen : 32,6 g, SD = 10,2 g ; PM moyenne : 24 %, SD = 0,04) (Figure S7). Les 32 CpG étaient associés à un effet indirect total de -55,36 g (SD = 51,28) sur le poids à la naissance.

Les CpGs ayant les effets indirects négatifs les plus importants incluent, cg10624729 (*P*-valeur = 5,15 x 10^{-8}), situé sur le gène *MIGA1* (Mitoguardin 1), associé à une diminution de 41,19 g du poids de naissance, cg19406975 (*P*-valeur = 9,27 x 10^{-8}), situé sur *SH3BP5L* (SH3 Binding Domain Protein 5 Like), un facteur d'échange de guanine, associé à une diminution de 39,97 g, cg01686933 (*P*-valeur = 6,98 x 10^{-7}), sur *NECTIN1* (Nectin Cell Molécule d'adhésion 1) qui code pour une protéine d'adhésion qui joue un rôle dans l'organisation des cellules épithéliales et endothéliales, associé à -40,67g, et cg14502606 (*P*-valeur = 1,04 x 10^{-6}), sur *MLX* (MAX Dimerization Protein MLX) un facteur de transcription qui joue un rôle dans la prolifération, la détermination et la différenciation des cellules, associé à -37,41 g (Figure 3.14, Tableau 3.2).

Tableau 3.2 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel sur le poids de naissance via les CpGs (FDR < 10%).

CpG			P-value		Max2		Indirect Effect		
Id	Position	Gene (distance)	X->M	M->Y	P-value	Q-value	Beta	CI	P-value
cg00025663	2:10827685	NOL10	0.0001884	0.00141008	7.58E-06	0.09178	-26.820	[-47.4; -6.23]	0.020175
cg00108098	3:122637664	SEMA5B	0.0001908	0.00073372	2.34E-06	0.04651	-28.680	[-55.65; -8.13]	0.000412
cg00232059	22:45131501	PRR5	0.000348	0.00154326	8.92E-06	0.09817	26.549	[6.56; 49.49]	0.000241
cg00692462	1:4773089	AJAP1	0.0003606	5.86E-05	4.66E-07	0.02353	-32.184	[-58.99; -10.49]	0.000322
cg00840341	16:17609453	XYLT1 (44714)	8.98E-06	0.00030177	4.73E-07	0.02366	-36.130	[-62.78; -13.56]	0.000813
cg01177854	16:68270387	ESRP2 (250)	0.0006578	0.0004485	1.43E-06	0.03368	26.983	[7.62; 45.53]	0.000681
cg01686933	11:119596104	NECTIN1	1.10E-06	0.00037453	6.98E-07	0.02734	-41.283	[-77.01; -12.94]	0.000493
cg02034222	2:74753281	DQX1	0.0003776	0.00102755	4.29E-06	0.06883	-27.698	[-54.67; -7.25]	0.000721
cg02151101	4:53439190	USP46 (17935)	2.24E-07	4.62E-05	1.58E-08	0.00298	47.514	[19.35; 82.63]	0.000145
cg07156115	7:17379753	AHR	4.44E-06	0.00038474	7.32E-07	0.02777	-39.427	[-69.65; -13.97]	0.000513
cg08694699	16:68270282	ESRP2 (145)	0.0005934	0.00024888	1.16E-06	0.03140	30.377	[8.55; 54.1]	0.000710
cg10406538	17:6024454	WSCD1	0.0015274	0.00038081	6.20E-06	0.08369	26.219	[9.73; 51.5]	0.000298
cg10624729	1:78298207	MIGA1	7.00E-06	8.86E-05	5.15E-08	0.00648	-41.303	[-74.5; -16.7]	0.000878
cg10778780	3:42307866	CCK (203)	0.0002953	0.00143873	7.86E-06	0.09321	26.300	[6.96; 50.18]	0.000792
cg11362604	15:37389585	MEIS2	0.0006456	5.28E-05	1.28E-06	0.03208	-31.162	[-54.46; -10.27]	0.000504
cg12160741	14:105760556	BRF1	0.0008164	0.0004502	2.08E-06	0.04331	-26.028	[-55.99; -6.73]	0.000761
cg12255774	11:61689144	RAB3IL1 (1402)	9.58E-05	0.00040328	7.97E-07	0.02850	-31.304	[-56.8; -12.7]	0.000178
cg13195895	1:147012788	BCL9 (392)	0.0017335	0.00097124	8.22E-06	0.09500	-24.633	[-46.09; -5.39]	0.020379
cg13800049	17:7466856	SENP3-EIF4A1	0.0003201	9.03E-05	3.72E-07	0.02126	-32.066	[-54.68; -14.07]	0.000775
cg14329026	20:35274941	SLA2 (321)	0.0004729	0.00111653	4.98E-06	0.07481	-24.767	[-48.77; -5.81]	0.000459
cg14502606	17:40718952	MLX (124)	0.0005533	3.57E-06	1.04E-06	0.03061	-38.472	[-67.62; -17.17]	0.000110
cg15035421	3:182631849	ATP11B	0.0006501	0.00152216	8.70E-06	0.09721	23.535	[7.21; 44.13]	0.020330
cg15676719	1:167598521	RCSD1 (951)	0.0010649	0.00013783	3.41E-06	0.05985	-31.054	[-60.24; -10.44]	0.020594
cg16593917	1:45793300	HPDL	0.0017779	0.001405	9.11E-06	0.09901	23.310	[4.32; 45.28]	0.020998
cg18817444	17:39306458	KRTAP4-5 (403)	5.46E-08	0.00120154	5.69E-06	0.08015	41.600	[16.15; 69.99]	0.000976
cg19406975	1:249110746	SH3BP5L	7.45E-06	0.00012249	9.27E-08	0.00874	-40.648	[-78.37; -14.48]	0.000589
cg20482145	17:8525165	MYH10	0.0017875	0.00011434	9.03E-06	0.09868	-28.116	[-48.13; -6.56]	0.000937
cg22211672	15:74610366	CCDC33	0.0006149	0.00013841	1.27E-06	0.03204	30.592	[10.89; 58.43]	0.000602
cg24571086	10:123371156	FGFR2 (13183)	0.0009816	0.00022289	2.99E-06	0.05499	-29.231	[-51.56; -10.84]	0.000335
cg25175240	15:69849786	DRAIC	0.0001818	0.00062392	1.75E-06	0.03870	32.070	[10.93; 55.65]	0.000849
cg26039141	11:75113160	RPS3	0.0004928	0.00042675	8.83E-07	0.02935	-28.224	[-50.67; -11.71]	0.000272
cg26814650	1:1269399	TAS1R3	1.75E-08	3.73E-05	1.07E-08	0.00298	55.920	[29.57; 86.86]	0.000687

À un niveau de FDR < 20 %, 164 médiateurs ont été découverts, dont 55 CpGs situés dans des régions enhancers et 26 CpG dans les régions promotrices. En comparaison avec l'ensemble du méthylome, il y a un enrichissement en régions enhancers (33% de tous les hits, *P* = 0,0003, test de Fisher), et il y a un appauvrissement en régions promotrices (15% de tous les résultats, *P* = 0,04, test de Fisher). 109 médiateurs appartenaient au corps d'un gène, et certains gènes ont été détectés plus d'une fois (*AJAP1*, *ESRP2*, *SH3BP2*, *SKI*, *SRSF5*, *VAV2* et *MLX*).

Concernant les deux CpGs, cg27402634 (entre *LINC00086* et *LEKR1*) et cg25585967 (*TRIO*), identifiés comme médiateurs dans (Morales et al., 2016), et pour le CpG, cg11280108 (*DCAF1*), présent dans la puce "HumanMethylation450 BeadChip" qui a été

identifié dans (Cardenas et al., 2019). Les associations entre le tabagisme et l'ADNm ont été retrouvées significatives pour ces trois CpGs (*P*-valeur = 9,07 x 10^{-14}) mais aucun de ces marqueurs ne sont identifié comme médiateurs dans nos travaux (*Q*-valeurs HDMAX2 > 0,93).

L'analyse par AMR a permis de détecter vingt-huit régions médiatrices de la relation tabagisme maternel et poids de naissance mais uniquement dix-neuf ont des effets indirects significatifs. Quatre sont situées dans des régions enhancer, sept dans des régions promotrices et vingt dans le corps d'un gène (Figure 3.13). Les dix-neuf AMRs sont associées à un effet indirect total de -52 g (SD = 45). L'effet indirect total des CpGs et AMRs est de -44 g (SD = 60, Tableau 3.3).

Tableau 3.3 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel sur le poids de naissance via les AMRs (P-valeur < 0.05).

AMR (chr:start-end	Gene (distance)	Nb CpGs	Comb-p		Indirect Effect		
			P-value	Q-value	Beta	CI	P-value
1 : 2171078 - 2171376	SKI	2	1.34E-06	2.87E-06	-18.85	[-39.29; -2.07]	0.04090
1 : 226926913 - 226927091	ITPKB (68)	3	7.51E-08	4.06E-07	12.54	[1.2; 31.92]	0.02028
10 : 122708861 - 122709152	WDR11 (39859)	2	2.79E-06	4.67E-06	17.33	[1.44; 40.37]	0.02071
10 : 8094553 - 8094802	GATA3-AS1	2	7.26E-10	1.34E-08	-16.95	[-34.24; -2]	0.02060
11 : 32355446 - 32355632	WT1 (53811)	2	4.32E-07	1.37E-06	-15.04	[-35.32; -1.12]	0.02055
16 : 433561 - 433838	LOC100134368	3	1.36E-06	2.87E-06	11.11	[0.65; 27.05]	0.02058
16 : 4714769 - 4714938	MGRN1	2	1.03E-06	2.38E-06	-22.08	[-44.12; -3.37]	0.00024
16 : 68269417 - 68270510	ESRP2	8	4.04E-18	1.86E-16	33.04	[12.75; 54.64]	0.00078
17 : 40713862 - 40715404	COASY	16	3.20E-13	9.82E-12	-25.89	[-44.06; -8.04]	0.00091
17 : 40718932 - 40719777	MLX	9	9.37E-19	8.62E-17	-26.66	[-50.89; -4.84]	0.00093
17 : 45949799 - 45950001	SP6 (16587)	2	9.13E-07	2.21E-06	-16.65	[-36.55; -4.39]	0.02044
17 : 79634966 - 79635196	CCDC137	2	2.86E-07	1.05E-06	-18.30	[-33.96; -4.27]	0.00077
20 : 36148579 - 36149235	BLCAP	29	9.84E-11	2.26E-09	-18.84	[-34.96; -5.23]	0.00030
3 : 195849490 - 195849797	LINC00885 (19831)	2	6.65E-07	1.89E-06	-25.82	[-54.64; -7.82]	0.00065
3 : 64211113 - 64211587	PRICKLE2 (23)	3	3.55E-08	2.72E-07	17.51	[2.33; 41.62]	0.02089
4 : 188916790 - 188917066	ZFP42	5	1.37E-06	2.87E-06	-11.80	[-31.77; -1.19]	0.00015
5 : 127872049 - 127872452	FBN2	3	1.38E-08	1.59E-07	20.84	[6.37; 45.74]	0.00062
6 : 2697880 - 2698160	MYLK4	3	8.53E-08	4.36E-07	-19.03	[-41.54; -3.64]	0.00034
7 : 130080968 - 130081362	CEP41	8	7.44E-06	9.78E-06	10.09	[0.67; 29.03]	0.04081

Douze AMRs sont associés à une diminution de poids de naissance (ACME moyen : -19.7 g, SD = 4,6 g ; PM moyenne : 15 %, SD = 4 %), et sept AMRs sont associés à une augmentation du poids de naissance (ACME moyen : 17,5 g, SD = 7,9 ; PM moyenne :

13 %, SD = 7 %) (Figure 3C). Le plus important effet négatif correspondait à l'AMR chr17:40 713 862-40 715 404 (*P* ajusté = 3,20 x 10 $^{-13}$) située sur *COASY* (Coenzyme A Synthase) qui joue un rôle important dans de nombreuses voies métaboliques de synthèse et de dégradation dans tous les organismes et associé à une diminution de 28,1 g. Cette AMR est située à seulement 3 Kb d'une autre AMR, chr17:40 718 932-40 719 777 (*P* ajusté = 9,37 x 10 $^{-19}$), située sur *MLX* , qui est associée à une diminution de 26,80 g du poids de naissance (Figure 3.15).

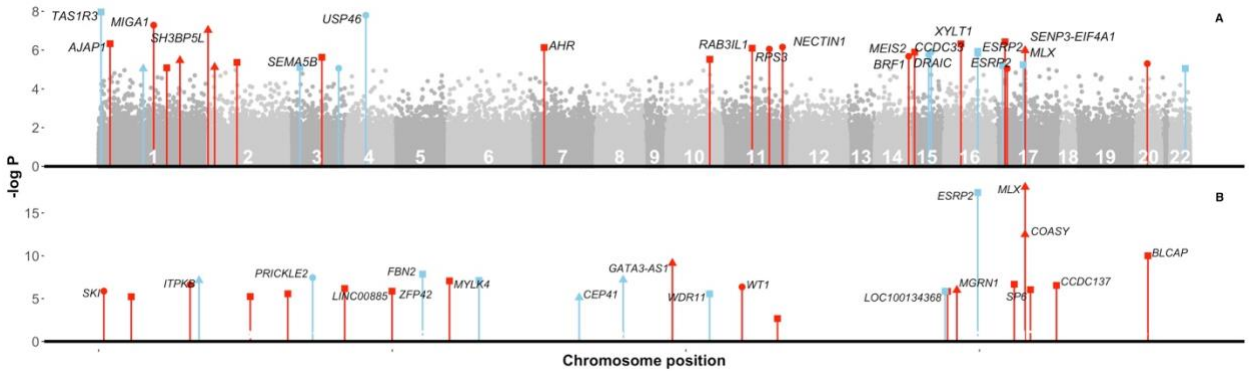

Figure 3.13 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel sur le poids de naissance. *A) Manhattan plot des -log (P-valeurs max-carré) obtenu à partir de HDMAX2 au niveau FDR de 10 % (CpG). Les noms de gènes correspondent aux hits respectant un niveau FDR de 5 % (32 hits). Les barres grises sans points correspondent aux CpG ne respectant pas un FDR de 10 %. B) Manhattan plot des -log10 (P-valeurs comb-p) pour les 28 AMRs. Les couleurs correspondent à l'intensité des effets indirects. Les symboles au-dessus des barres colorées correspondent à la catégorisation en enhancer, promoteur ou inconnu.*

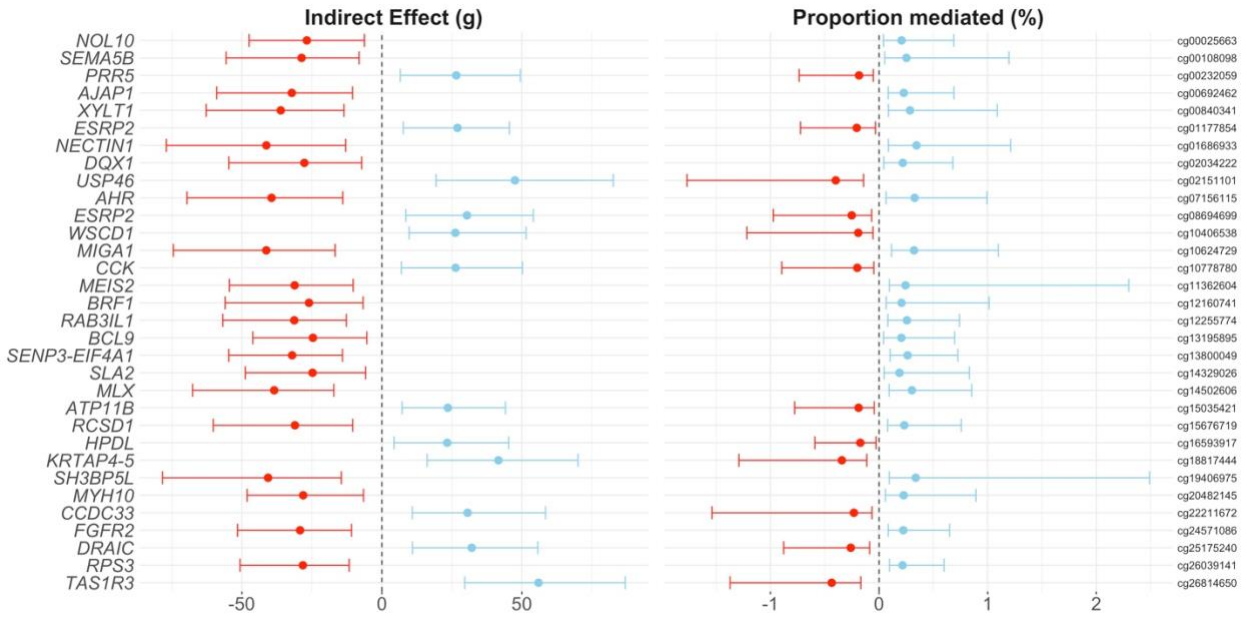

Figure 3.14 : Estimations des effets indirects (ACME) et proportions médiées pour les 32 CpGs médiant la relation du tabagisme maternel sur le poids de naissance. *Les estimations des effets indirects et de leurs intervalles de confiance ont été calculées à l'aide du package "mediation" (10000 simulations de MCMC).*

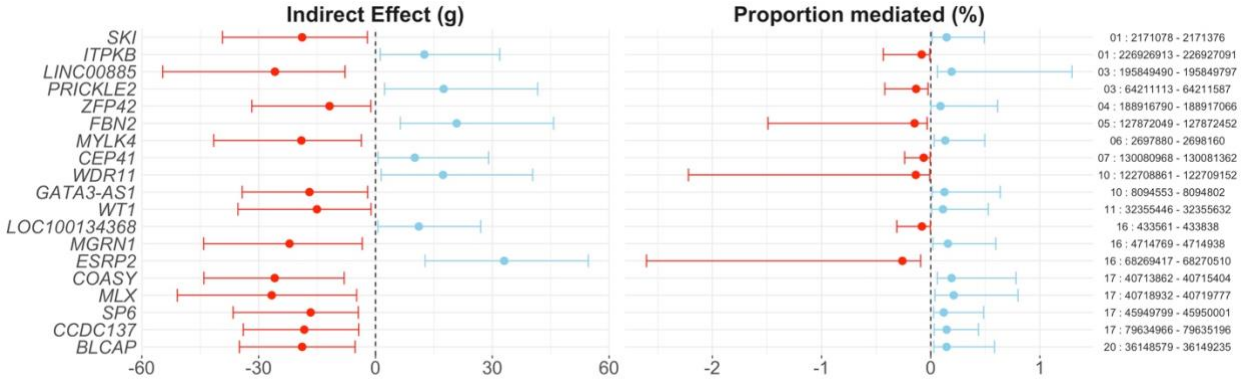

Figure 3.15 : Estimations des effets indirects (ACME) et proportions médiées pour les 19 AMRs médiant la relation du tabagisme maternel sur le poids de naissance. *Les estimations des effets indirects et de leurs intervalles de confiance ont été calculées à l'aide du package "mediation" (10000 simulations de MCMC).*

Comparaison des résultats de HDMAX2

Nous avons également réalisé l'analyse de médiation tabagisme maternel – méthylation – poids à la naissance avec différentes combinaisons de méthodes. Pour les EWAS (étape 1), utilisation soit de LFMM soit de RefFreeEWAS et pour le test de médiation (étape 2), utilisation du test max2 ou de HDMT. HDMAX2 a été comparé à : LFMM-HDMT, RefFreeEWAS-max2 et RefFreeEWAS-HDMT.

Si l'on regarde les résultats, concernant la comparaison de HDMAX2 avec LFMM hdmt (Figure 3.16A), les méthodes ordonnent de la même manière les *P*-valeurs mais HDMAX2 permet de détecter, sous un seuil FDR de 10%, plus de CpGs (32 contre 25). On retrouve le même type de résultats pour la comparaison de RefFreeEWAS-max2 et RefFreeEWAS-hdmt, c'est-à-dire qu'on a exactement le même ordre de *P*-valeurs mais plus de découverte avec le test de médiation max2 (20 contre 5). Ces résultats sont confirmés par le tableau 3.4 qui nous montre bien que les rangs des premiers hits sont similaires entre HDMAX2 et LFMM hdmt et entre RefFreeEWAS-max2 et RefFreeEWAS-hdmt.

Néanmoins, si on compare des méthodes d'EWAS (Figure 3.16BCDE), on retrouve des différences dans les ordres de *P*-valeurs mais tout de même un certain nombre de CpGs sont toujours découverts quel que soit la méthode testée, tel que : USP46, SH3BP5L et MEIS2. Ceci est confirmé dans le tableau 3.4 qui permet de voir que les rangs des CpGs sont différents entre les méthodes utilisant LFMM et RefFreeEWAS.

Pour conclure sur cette comparaison de méthodes sur l'analyse tabagisme – méthylation – poids de naissance, on a pu constater que les deux tests de médiation (étape 2) max2 et HDMT, qui montrent les meilleures performances dans les études de simulations, donnent les mêmes résultats sur des données réelles, à la différence près que max2 permet de détecter davantage de CpGs, mais nous avons montré via le calcul du eta0 que max2 respectait bien le FDR. Enfin, max2 est beaucoup plus rapide que HDMT (environ 3 secondes contre plus d'une heure). Concernant l'étape 1, les simulations avaient montré que dans la majorité des scénarios, LFMM était supérieure à RefFreeEWAS. Sur les données réelles nous n'avons pas de résultat de référence, il est donc délicat de comparer les résultats obtenus par LFMM et par RefFreeEWAS. Cependant, on constate que le Eta0 de RefFreeEWAS-max2 était égal à 1 tandis que celui de HDMAX2 est égal à 0.998974. Il a été démontré qu'une valeur d'Eta0 légèrement inférieure à 1 permettait d'avoir des tests statistiques plus puissants tandis qu'un Eta0 égal à 1 engendre des résultats conservatifs (Benjamini and Hochberg, 2000, 1995; Storey, 2002).

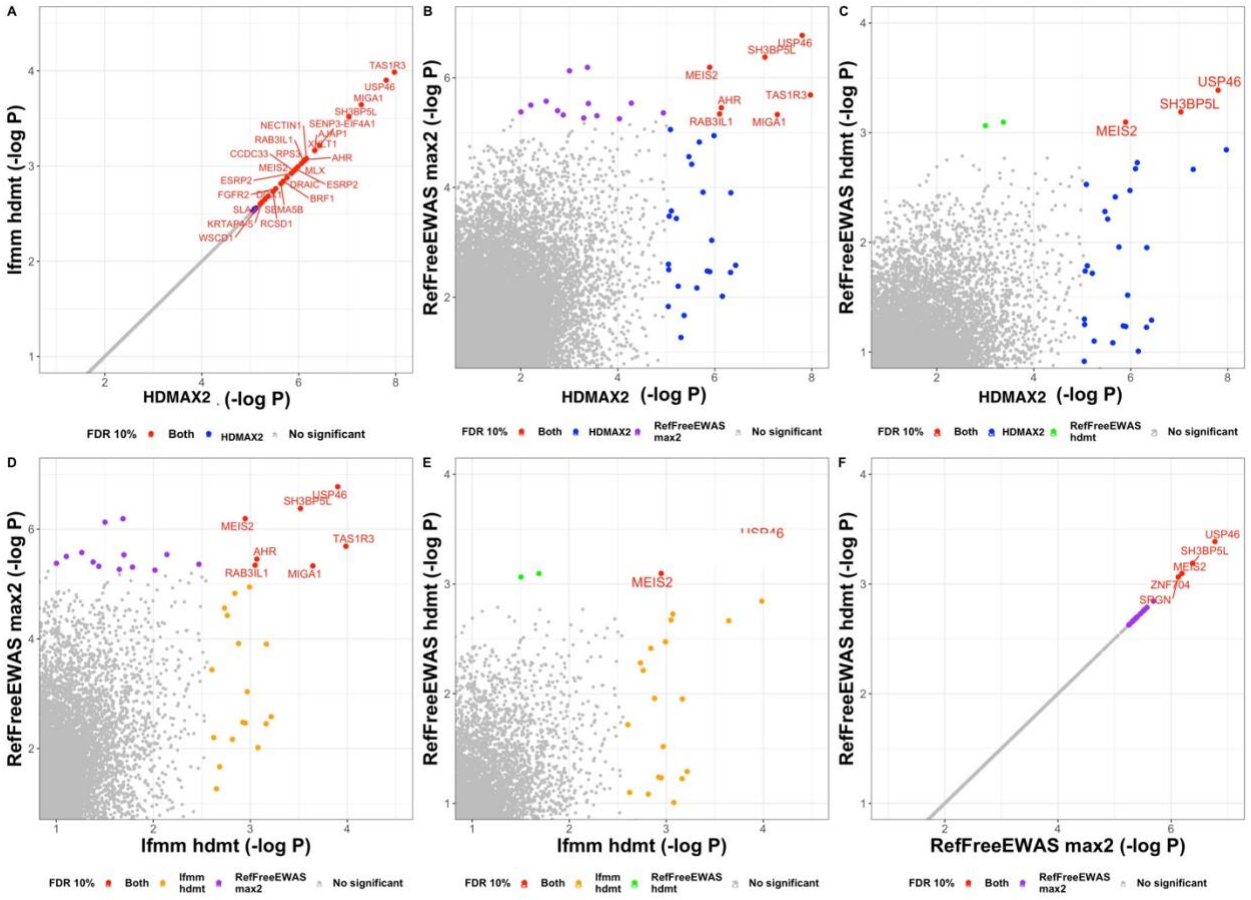

Figure 3.16 : Comparaison des P-valeurs entre méthodes de médiation en haute dimension sur l'analyse Tabac – Méthylation – Poids de naissance. *HDMAX2 correspond à la combinaison des méthodes LFMM et max2. Les points de couleurs représentent les CpGs respectant un seuil FDR de 10%. Les points rouges correspondent aux CpGs détectés par les deux méthodes de chaque panel.*

Tableau 3.4 : Les 15 premiers CpGs de HDMAX2 sur l'analyse Tabac – Méthylation – Poids de naissance et les rangs respectifs des autres méthodes testées.

CpG	Gene	Rank				P-Values			
		lfmm max2	lfmm hdmt	RefFreeEWAS max2	REfFreeEWAS hdmt	lfmm max2	lfmm hdmt	RefFreeEWAS max2	RefFreeEWAS hdmt
cg26814650	TAS1R3	1	1	6	6	1.07E-08	0.00010	2.06E-06	0.00143
cg02151101	USP46	2	2	1	1	1.58E-08	0.00013	1.69E-07	0.00041
cg10624729	MIGA1	3	3	16	16	5.15E-08	0.00023	4.68E-06	0.00216
cg19406975	SH3BP5L	4	4	2	2	9.27E-08	0.00030	4.20E-07	0.00065
cg13800049	SENP3-EIF4A1	5	5	2833	2833	3.72E-07	0.00061	2.63E-03	0.05126
cg00692462	AJAP1	6	6	326	326	4.66E-07	0.00068	1.25E-04	0.01118
cg00840341	XYLT1	7	7	3453	3453	4.73E-07	0.00069	3.54E-03	0.05951
cg01686933	NECTIN1	8	8	6892	6892	6.98E-07	0.00084	9.67E-03	0.09832
cg07156115	AHR	9	9	11	11	7.32E-07	0.00086	3.52E-06	0.00188
cg12255774	RAB3IL1	10	10	15	15	7.97E-07	0.00089	4.55E-06	0.00213
cg26039141	RPS3	11	11	246300	246300	8.83E-07	0.00094	6.49E-01	0.80557
cg14502606	MLX	12	12	35	35	1.04E-06	0.00102	1.14E-05	0.00337
cg08694699	ESRP2	13	13	1387	1387	1.16E-06	0.00108	9.21E-04	0.03035
cg22211672	CCDC33	14	14	3374	3374	1.27E-06	0.00113	3.43E-03	0.05855
cg11362604	MEIS2	15	15	3	3	1.28E-06	0.00113	6.43E-07	0.00080

3.4.2.2. Analyse de l'âge gestationnel

Une seconde analyse de médiation en haute dimension a été réalisée pour comprendre l'impact du tabagisme maternel sur l'âge gestationnel. Pour un FDR de 10 %, quinze CpGs (2 CpGs à un niveau FDR < 5 %) sont identifiés comme médiateurs (figure 3.17, *P* max-carré < 3,27 x 10^{-6}). Les 15 CpG sont associés à un effet indirect total négatif de -0,196 semaine de grossesse (sd = 0,11).

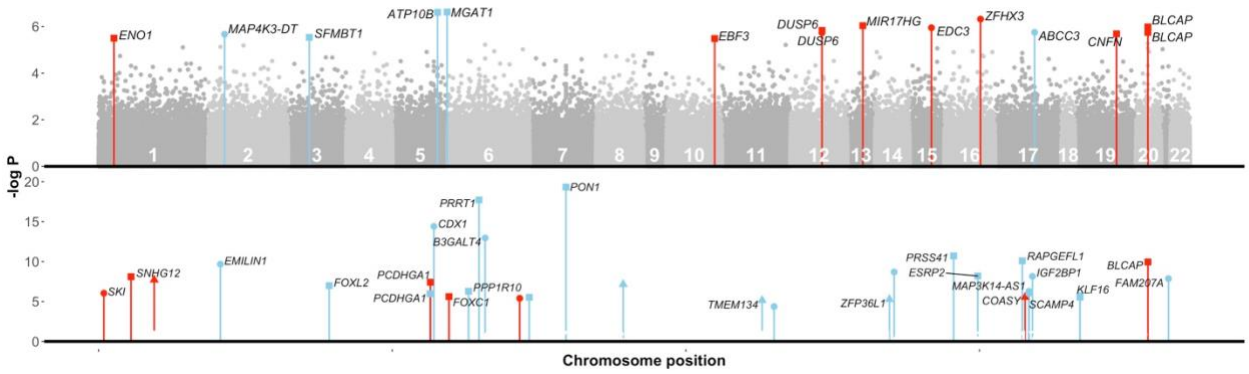

Figure 3.17 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel sur le l'âge gestationnel. *A) Manhattan plot des -log (P-valeurs max-carré) obtenu à partir de HDMAX2 au niveau FDR de 20 % (CpG). Les noms de gènes correspondent aux hits respectant un niveau FDR de 10 % (32 hits). Les barres grises sans points correspondent aux CpG ne respectant pas un FDR de 20 %. B) Manhattan plot des -log10 (P-valeurs comb-p) pour les 28 AMRs. Les couleurs correspondent à l'intensité des effets indirects. Les symboles au-dessus des barres colorées correspondent à la catégorisation en enhancer, promoteur ou inconnu.*

Les CpGs ayant les effets indirects les plus importants incluent, cg10298741 (*P*-valeur = 4,81 x 10^{-7}), situé sur le gène *ZFHX3* (Zinc Finger Homeobox 3), un facteur de transcription qui régule la différenciation myogénique et neuronale, associé à une diminution de l'âge gestationnel de 0,046 semaine, cg04908961 (*P*-valeur = 9,18 x 10^{-7}), sur *MIR17HG* (MiR-17-92a-1 Cluster Host Gene), un gène hôte pour le cluster MIR17-

92, un groupe de microARN (miARN) impliqués dans la survie, la prolifération et la différenciation cellulaire, associée à une diminution de l'âge gestationnel de 0,046 semaine, cg08402058 (*P*-valeur = 1,04 x 10^{-6}), sur *BLCAP* (Bladder Cancer Associated Protein) qui réduit la croissance cellulaire en stimulant l'apoptose, associée à une diminution de l'âge gestationnel de 0,040 semaine, (Figure 3.18, Tableau 3.5).

Tableau 3.5 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel l'âge gestationnel via les CpGs (FDR < 10%).

CpG			P-value		Max2		Indirect Effect		
Id	Position	Gene (distance)	X->M	M->Y	P-value	Q-value	Beta	CI	P-value
cg04671476	5:180240653	MGAT1	0.00024848	8.28E-05	2.38E-07	0.04642	0.0980	[0.03; 0.17]	0.000705
cg04908961	13:92003278	MIR17HG	0.0005335	2.50E-05	9.19E-07	0.06274	-0.0922	[-0.18; -0.03]	0.000845
cg06344553	10:131684893	EBF3	0.00054205	0.00104013	3.28E-06	0.08266	-0.0757	[-0.15; -0.02]	0.000294
cg08402058	20:36148961	BLCAP	0.00050089	0.000559	1.04E-06	0.06370	-0.0856	[-0.16; -0.03]	0.000736
cg10298741	16:73019173	ZFHX3	0.00037052	0.00028401	4.82E-07	0.05621	-0.0845	[-0.15; -0.03]	0.000589
cg10655499	3:53084209	SFMBT1 (4119)	0.0009681	0.00062104	2.93E-06	0.07875	0.0812	[0.02; 0.14]	0.000387
cg13432294	1:8937640	ENO1	0.00013769	0.00102251	3.18E-06	0.08157	-0.0853	[-0.14; -0.03]	0.000678
cg14327995	19:42894499	CNFN (54)	0.00042127	0.00081025	2.07E-06	0.06755	-0.0787	[-0.14; -0.03]	0.000999
cg14389122	15:74945851	EDC3	0.0001022	0.00058442	1.13E-06	0.06429	-0.0900	[-0.17; -0.04]	0.000484
cg17740822	12:89744609	DUSP6	0.00067836	0.00028502	1.47E-06	0.06592	-0.0867	[-0.15; -0.03]	0.000470
cg18426477	17:48764504	ABCC3	0.00067889	0.00074713	1.78E-06	0.06689	0.0671	[0.02; 0.13]	0.000122
cg19548448	5:160278962	ATP10B	8.71E-05	0.000256	2.45E-07	0.04642	0.0957	[0.03; 0.19]	0.000966
cg20722088	12:89742886	DUSP6	4.34E-10	0.00076439	1.86E-06	0.06709	-0.1351	[-0.21; -0.04]	0.000216
cg20783699	20:36148767	BLCAP	0.00076085	0.0004242	1.80E-06	0.06693	-0.0788	[-0.16; -0.02]	0.000753
cg21101086	2:39836209	MAP4K3-DT (4992)	0.00081174	5.64E-05	2.15E-06	0.06771	0.0860	[0.04; 0.15]	0.000214

Dix CpGs sont associés à une diminution de l'âge gestationnel (ACME moyen = -0,045 semaine, sd = 0,0092 ; PM : 43 %, sd = 6 %) et cinq CpGs sont associés à une augmentation de l'âge gestationnel (ACME moyen = 0,043 semaine, sd = 0,0052 ; PM : 39 %, SD = 9 %).

Pour un niveau de FDR de 20 %, soixante-trois médiateurs sont identifiés, dont vingt-six CpGs dans des régions enhancers. Ceci correspond à un enrichissement significatif en en région enhancer (41 % de tous les CpGs, $P < 2{,}2 \times 10^{-16}$, test de Fisher).

L'analyse par AMRs a abouti à la détection de trente et un médiateurs dont vingt-trois ayant des effets indirects significatifs. Onze AMRs sous situées dans des régions

enhancers, sept dans les régions promotrices et vingt-six dans le corps d'un gène. Les vingt-trois AMRs sont associées à un effet indirect total de 0,12 semaine entraînant un âge gestationnel plus long (sd = 0,11). L'effet indirect total cumulé des CpGs et des AMRs est de -0,09 semaine (sd = 0,14). Les vingt-trois AMRs sont associées à des effets indirects faibles mais statistiquement significatifs allant de -0,09 semaine à 0,010 semaine (aucun n'est associé à une proportion médiée significative). Cinq AMRs sont associées à une diminution de l'âge gestationnel (ACME moyen = -0,0649 semaine, sd = 0,0156), et dix-huit AMRs sont associées à un âge gestationnel plus long (ACME moyen = 0.0593 semaine, sd = 0,0134).

Les effets indirects les plus importants correspondent aux AMR suivant: chr1 : 28 906 332 - 28 906 661 (*P* ajusté = 7,89 x 10^{-9}) situé sur le gène *SNHG12* (Small Nucleolar RNA Host Gene 12), un gène qui peut favoriser la tumorigenèse, associé à un âge gestationnel plus court de 0,041 semaine, chr20 : 36 148 579 - 36 149 354 (*P* ajusté = 1,13 x 10^{-10}) sur *BLCAP*, associée à une diminution de l'âge gestationnel 0,031 semaine, et chr17:40 714 100 - 40 714 374 (*P* ajusté = 2,84 x 10^{-6}) sur *COASY* associé à une diminution de l'âge gestationnel de 0,027 semaine (Figure 3.19, Tableau 3.6).

Tableau 3.6 : Analyse de médiation en haute dimension (HDMAX2) du tabagisme maternel l'âge gestationnel via les AMRs (P-valeur < 0.05).

AMR (chr:start-end	Gene (distance)	Nb CpGs	Comb-p		Indirect Effect		
			P-value	Q-value	Beta	CI	P-value
1 : 28906332 - 28906661	SNHG12	4	7.89E-09	4.41E-08	-0.0921	[-0.15; -0.03]	0.00026
11 : 67232391 - 67232634	TMEM134	2	7.13E-06	9.56E-06	0.0514	[0.01; 0.11]	0.04085
14 : 69256799 - 69257100	ZFP36L1	2	5.29E-06	7.38E-06	0.0424	[0; 0.09]	0.02014
16 : 2848919 - 2849267	PRSS41	2	1.94E-11	2.60E-10	0.0583	[0.02; 0.13]	0.00024
16 : 68270251 - 68270510	ESRP2 (115)	4	6.43E-09	4.31E-08	0.0700	[0.01; 0.14]	0.00079
17 : 38334055 - 38334562	RAPGEFL1	6	8.14E-11	9.09E-10	0.0661	[0.02; 0.12]	0.00038
17 : 40714100 - 40714374	COASY	2	2.84E-06	4.55E-06	-0.0600	[-0.12; -0.01]	0.02056
17 : 43339450 - 43339954	MAP3K14-AS1	8	5.35E-07	1.44E-06	0.0625	[0.01; 0.13]	0.04012
17 : 47091461 - 47092395	IGF2BP1	6	7.09E-09	4.32E-08	0.0592	[0.01; 0.12]	0.00046
19 : 1852004 - 1852288	KLF16 (231)	2	2.01E-06	3.64E-06	0.0493	[0; 0.12]	0.04082
19 : 1908254 - 1908511	SCAMP4	2	2.90E-06	4.55E-06	0.0588	[0.01; 0.13]	0.04067
2 : 27308999 - 27309276	EMILIN1	2	2.12E-10	1.77E-09	0.0703	[0.02; 0.14]	0.00070
20 : 36148579 - 36149354	BLCAP	29	1.13E-10	1.08E-09	-0.0620	[-0.12; -0.02]	0.00065
21 : 46378365 - 46378748	FAM207A	2	1.39E-08	7.15E-08	0.0522	[0.01; 0.12]	0.02022
3 : 138658365 - 138658677	FOXL2 (4510)	2	1.04E-07	3.88E-07	0.0411	[0; 0.09]	0.04039
5 : 140723577 - 140723807	PCDHGA1	2	1.07E-06	2.47E-06	0.0396	[0.01; 0.09]	0.02076
5 : 149546195 - 149547069	CDX1	8	3.98E-15	8.88E-14	0.0684	[0.01; 0.15]	0.00011
6 : 1608518 - 1608790	FOXC1 (2012)	2	2.44E-06	4.19E-06	-0.0516	[-0.1; -0.01]	0.00056
6 : 30582296 - 30582539	PPP1R10	2	5.37E-07	1.44E-06	0.0640	[0.02; 0.12]	0.00047
6 : 32116086 - 32117211	PRRT1	23	1.92E-18	6.42E-17	0.0524	[0.02; 0.11]	0.00019
6 : 33245303 - 33245927	B3GALT4	23	1.12E-13	1.87E-12	0.0963	[0.04; 0.17]	0.00081
7 : 95025733 - 95026625	PON1 (48)	18	4.74E-20	3.17E-18	0.0654	[0; 0.13]	0.02063

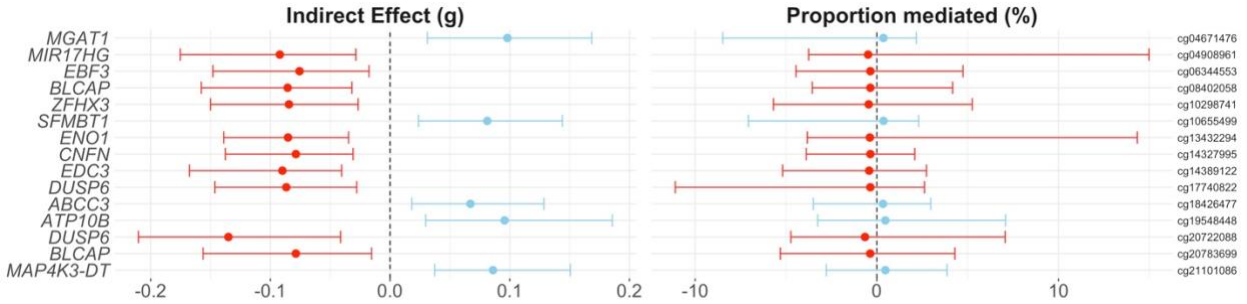

Figure 3.18 : Estimations des effets indirects (ACME) et proportions médiées pour les 15 CpGs médiant la relation du tabagisme maternel sur l'âge gestationnel. *Les estimations des effets indirects et de leurs intervalles de confiance ont été calculées à l'aide du package "mediation" (10000 simulations de MCMC).*

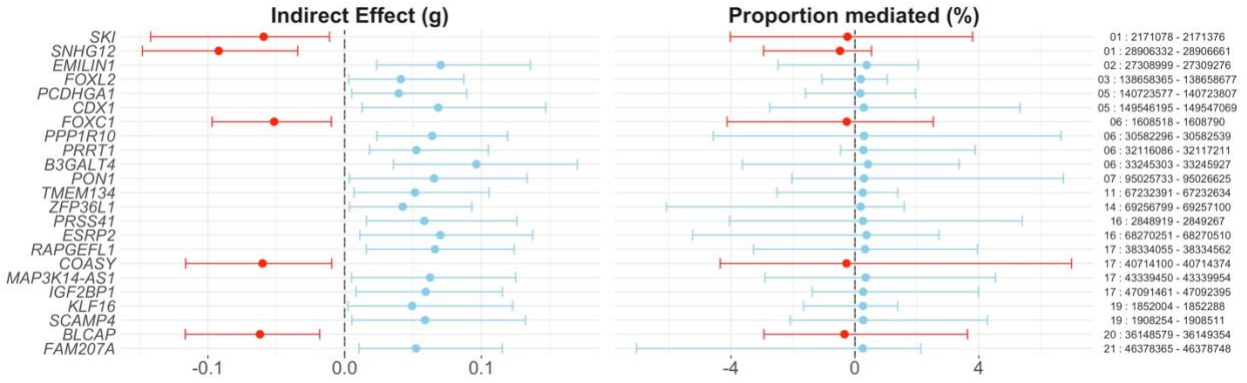

Figure 3.19 : Estimations des effets indirects (ACME) et proportions médiées pour les 23 AMRs médiant la relation du tabagisme maternel sur l'âge gestationnel. *Les estimations des effets indirects et de leurs intervalles de confiance ont été calculées à l'aide du package "mediation" (10000 simulations de MCMC).*

Comparaison des résultats de HDMAX2

Nous avons également réalisé l'analyse de médiation tabagisme maternel – méthylation – âge gestationnel avec différentes combinaisons de méthodes. Pour les EWAS (étape 1), soit LFMM soit RefFreeEWAS ont été utilisés et pour le test de médiation (étape 2), soit max2 ou soit HDMT ont été utilisés. HDMAX2 a été comparé à : LFMM-HDMT, RefFreeEWAS-max2 et RefFreeEWAS-HDMT.

Les méthodes HDMT et LFMM-hdmt (Figure 3.20A), ordonnent de la même manière les *P*-valeurs, permettant de détecter, sous un seuil FDR de 10%, autant de CpGs. On retrouve le même type de résultats pour la comparaison de RefFreeEWAS-max2 et RefFreeEWAS-hdmt, c'est-à-dire qu'on a exactement le même ordre de *P*-valeurs et que l'on détecte les mêmes CpGs. Ces résultats sont confirmés par le tableau 3.3 qui nous montre bien que les rangs des premiers hits sont similaires entre HDMAX2 et LFMM-hdmt et entre RefFreeEWAS-max2 et RefFreeEWAS-hdmt.

Toutefois si on compare les méthodes d'EWAS (Figure 3.20BCDE), on retrouve des différences dans les ordres de *P*-valeurs mais tout de même un certain nombre de CpGs sont toujours découverts quel que soit la méthode testée, tels que : MAP4K3-DT et ATP10B. Ceci est confirmé dans le tableau 3.3 qui permet de voir que les rangs des CpGs sont différents entre les méthodes utilisant LFMM et RefFreeEWAS.

Pour conclure sur cette comparaison de méthodes sur l'analyse tabagisme – méthylation – âge gestationnel, on a pu constater que les deux tests de médiation (étape 2) max2 et HDMT donnent les mêmes résultats sur ces données. Pour finir, comme il a été précédemment montré, max2 est beaucoup plus rapide que HDMT (environ 3 secondes contre plus d'une heure). Concernant l'étape 1, les simulations avaient montré que dans la majorité des scénarios LFMM était supérieure à RefFreeEWAS, mais n'ayant pas de référence sur les données réelles, il est délicat de comparer résultats obtenus par LFMM à ceux obtenus avec RefFreeEWAS. Néanmoins, nous avons constaté que le Eta0 de RefFreeEWAS-max2 était égal à 1 tandis que celui de HDMAX2 est égal à 0.9978999. Et donc comme pour l'analyse portant sur le poids de naissance, l'utilisation de LFMM permettrait de réaliser des tests statistiques plus puissants.

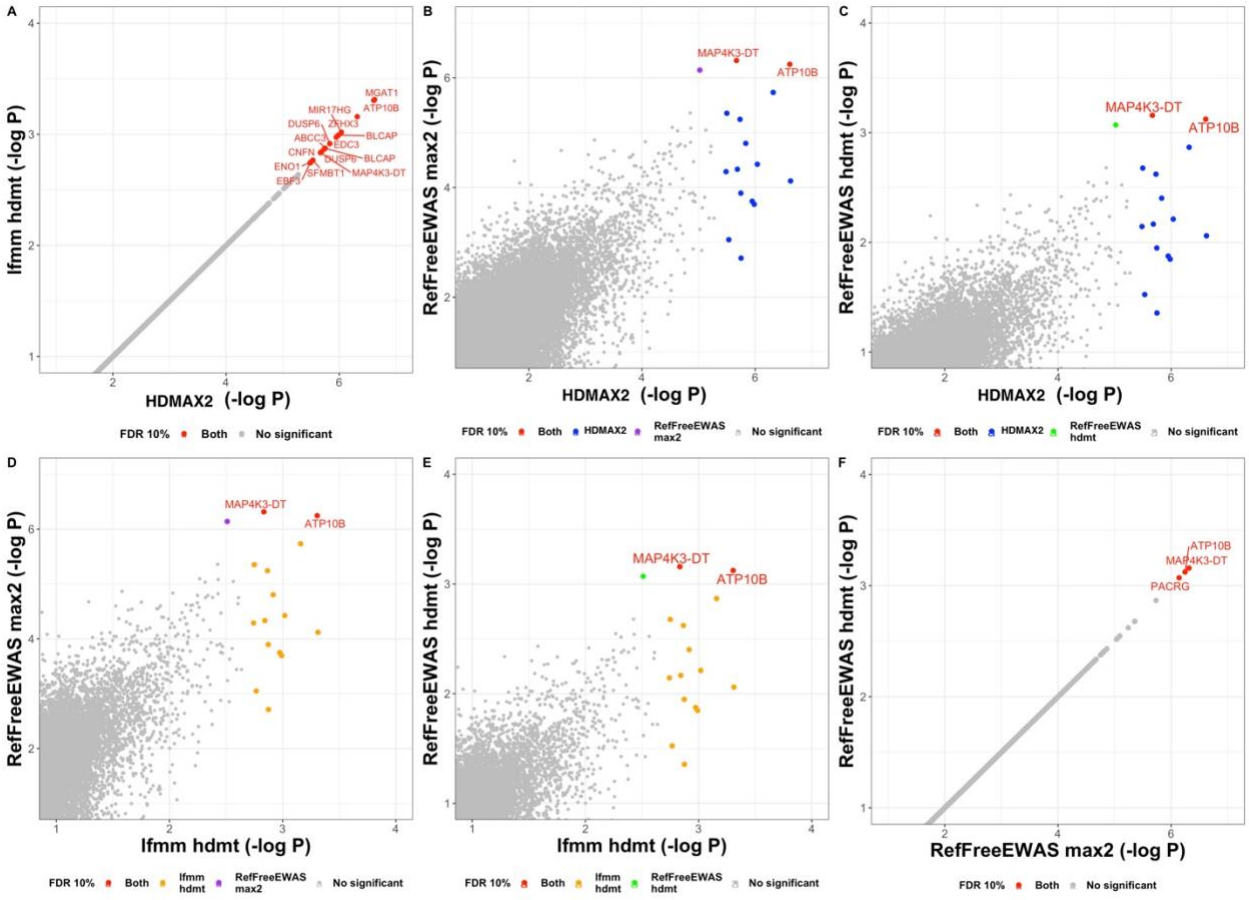

Figure 3.20 : Comparaison des P-valeurs entre méthodes de médiation en haute dimension sur l'analyse Tabac – Méthylation – âge gestationnel. *HDMAX2 correspond à la combinaison des méthodes LFMM et max2. Les points de couleurs représentent les CpGs respectant un seuil FDR de 10%. Les points rouges correspondent aux CpGs détectés par les deux méthodes de chaque panel.*

Tableau 3.3 : Les 15 premiers CpGs de HDMAX2 sur l'analyse Tabac – Méthylation – âge gestationnel et les rangs respectifs des autres méthodes testées.

CpG	Gene	Rank				P-Values			
		lfmm max2	lfmm hdmt	RefFreeEWAS max2	RefFreeEWAS hdmt	lfmm max2	lfmm hdmt	RefFreeEWAS max2	RefFreeEWAS hdmt
cg04671476	MGAT1	1	1	64	64	2.4E-07	4.9E-04	7.6E-05	8.7E-03
cg19548448	ATP10B	2	2	2	2	2.5E-07	5.0E-04	5.7E-07	7.5E-04
cg10298741	ZFHX3	3	3	4	4	4.8E-07	6.9E-04	1.8E-06	1.4E-03
cg04908961	MIR17HG	4	4	40	40	9.2E-07	9.6E-04	3.8E-05	6.1E-03
cg08402058	BLCAP	5	5	147	147	1.0E-06	1.0E-03	2.0E-04	1.4E-02
cg14389122	EDC3	6	6	140	140	1.1E-06	1.1E-03	1.8E-04	1.3E-02
cg17740822	DUSP6	7	7	15	15	1.5E-06	1.2E-03	1.6E-05	4.0E-03
cg18426477	ABCC3	8	8	1080	1080	1.8E-06	1.3E-03	1.9E-03	4.4E-02
cg20783699	BLCAP	9	9	106	106	1.8E-06	1.3E-03	1.3E-04	1.1E-02
cg20722088	DUSP6	10	10	7	7	1.9E-06	1.4E-03	5.7E-06	2.4E-03
cg14327995	CNFN	11	11	48	48	2.1E-06	1.4E-03	4.7E-05	6.8E-03
cg21101086	MAP4K3-DT	12	12	1	1	2.1E-06	1.5E-03	4.8E-07	7.0E-04
cg10655499	SFMBT1	13	13	560	560	2.9E-06	1.7E-03	9.0E-04	3.0E-02
cg13432294	ENO1	14	14	6	6	3.2E-06	1.8E-03	4.4E-06	2.1E-03
cg06344553	EBF3	15	15	50	50	3.3E-06	1.8E-03	5.2E-05	7.2E-03

3.4.2.3. Analyse des médiateurs communs entre le poids à la naissance et l'âge gestationnel

Les deux analyses de médiation en haute dimension ont mis en évidence la présence d'AMR communes entre la relation tabagisme-poids de naissance et la relation tabagisme-âge gestationnel. Ces AMRS sont situées sur six gènes, *COASY, BLCAP*, *SKI*, *DECR1*, *ESRP2, PRRT1*.

Pour mieux comprendre la causalité entre l'âge gestationnel et le poids de naissance, nous avons réalisé une nouvelle étude de médiation portant sur l'impact de l'âge gestationnel sur le poids à la naissance via les six AMRs en commun. Dans cette analyse, l'âge gestationnel a des effets indirects significatifs sur le poids à la naissance pour les AMRs situés sur *COASY* (ACME = 6.9 g, *P* médiation = 0,001) et *BLCAP* (ACME = 5,1 g, *P* médiation = 0,01) (Figure 3.21). Ces deux AMRs sont associées à un effet indirect total de 10 g sur le poids de naissance (SD = 3.9).

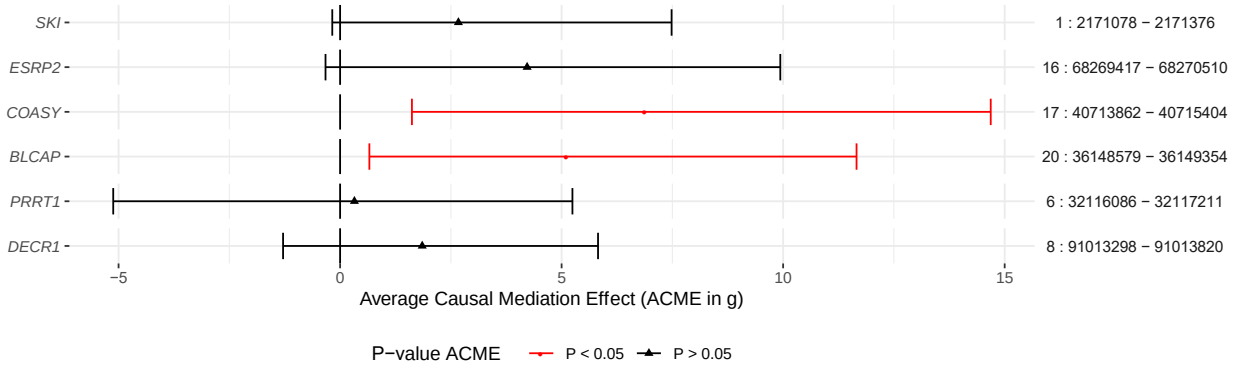

Figure 3.21 : Effets indirects (ACME) des 6 AMRs dans l'analyse de médiation de l'âge gestationnel sur le poids de naissance. *Les effets indirects et leurs intervalles de confiance à 95 % ont été estimés à l'aide du package "mediation". L'effet indirect total des deux AMR est de 10 g (sd = 3.9).*

Dans l’optique de compléter l’analyse de ces deux AMRs nous avons zoomé sur leurs localisations génétiques.

Concernant la zone du gène *COASY* (figure 3.22) on retrouve deux AMRs, l’un sur le gène *COASY* et l’autre sur le gène *MLX*. Les AMRs sont situées dans une région hypométhylée (figure 3.22B) et le tabagisme maternel induit une diminution des niveaux d'ADNm (figure 3.22C). Les effets indirects des CpGs contenus dans les AMRs induisent une diminution du poids à la naissance et figurent parmi les effets les plus négatifs (Figure 3.22DE).

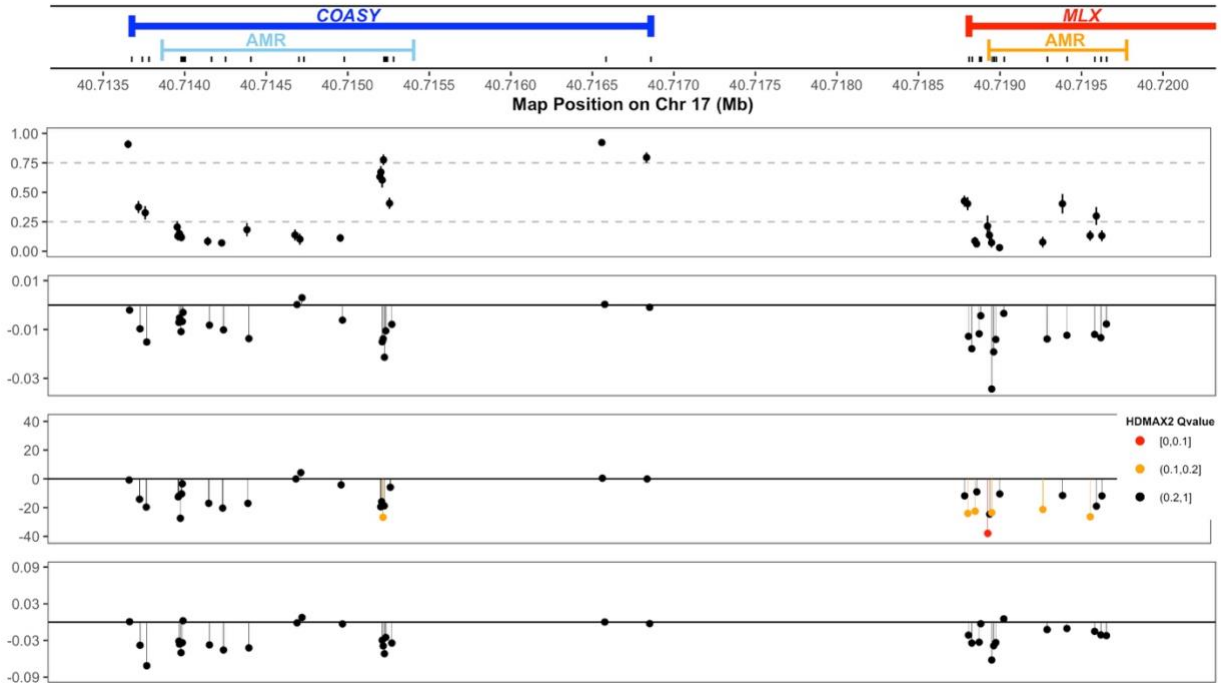

Figure 3.22 : AMRs et CpGs dans la région COASY - MLX du chromosome 17. *A) Position chromosomique, B) Moyenne et écart type des niveaux de méthylation pour les CpGs. C) Taille d'effets du tabagisme maternel sur les niveaux d'ADNm pour chaque CpG. Effets indirects dans la médiation du tabagisme maternel sur le poids de naissance pour chaque CpG. E) Effets indirects dans la médiation du tabagisme maternel sur l'âge gestationnel pour chaque CpG.*

Dans la région génomique entourant le gène *BLCAP* (Figure 3.23), les AMRs sont situées dans des zones hyperméthylées (Figure 3.23B) et le tabagisme maternel est lié à une diminution des niveaux d'ADNm dans les AMRs (Figure 3.23C). Les effets indirects des CpGs contenus dans l'AMR induisent une diminution du poids de naissance (Figure 3.23D).

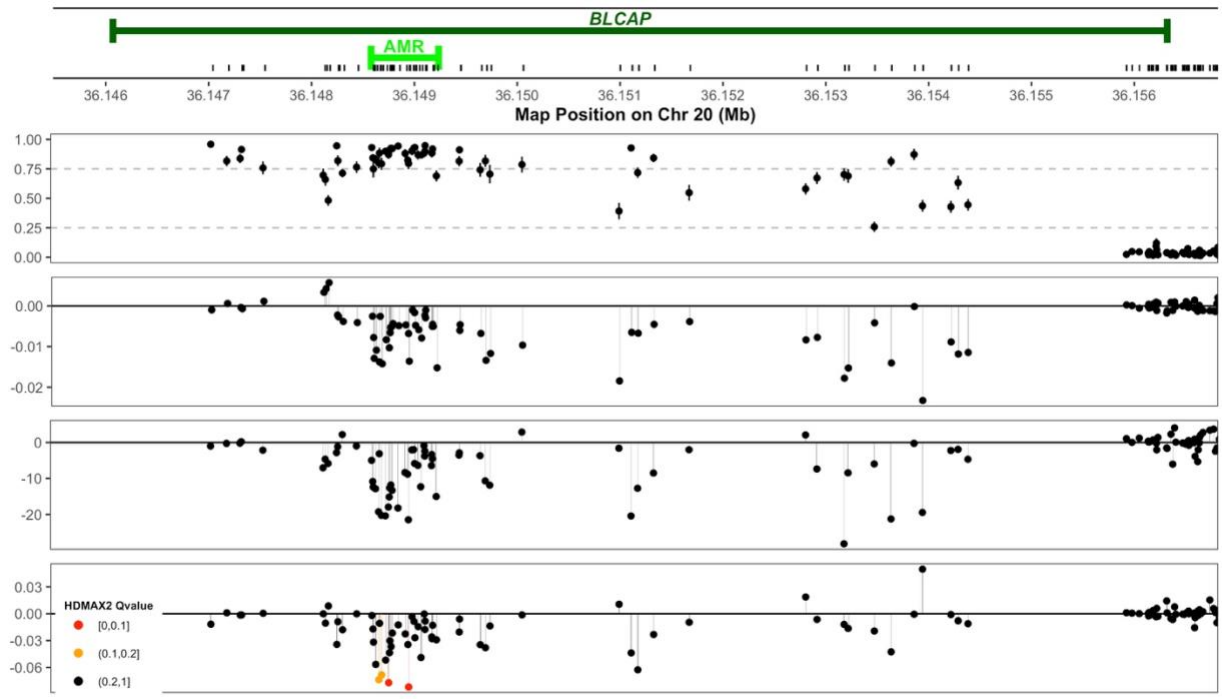

Figure 3.23 : AMRs et CpGs dans la région BLCAP du chromosome 20. *A) Position chromosomique, B) Moyenne et écart type des niveaux de méthylation pour les CpGs. C) Taille d'effets du tabagisme maternel sur les niveaux d'ADNm pour chaque CpG. Effets indirects dans la médiation du tabagisme maternel sur le poids de naissance pour chaque CpG. E) Effets indirects dans la médiation du tabagisme maternel sur l'âge gestationnel pour chaque CpG.*

3.4.3. Discussion

Après le développement de notre procédure de médiation en haute dimension et l'évaluation de ces performances, nous avons réalisé une analyse visant à comprendre les mécanismes de causalités entre le tabagisme maternel, la méthylation placentaire (ADNm) et la santé future de l'enfant. Comme biomarqueur de la santé future, nous nous sommes principalement intéressés au poids à la naissance à terme mais aussi à l'âge gestationnel à terme.

Nous avons donc évalué les effets indirects, de l'ADNm, de l'exposition du tabagisme maternel sur le poids de naissance mais aussi sur l'âge gestationnel.
Nous avons découvert que l'effet indirect total du méthylome entraîne une diminution d'environ 44 g du poids de naissance, représentant 31 % de l'effet total (-140 g). Par rapport aux analyses de médiation précédentes du tabagisme sur le poids de naissance (Cardenas et al., 2019; Morales et al., 2016; Xu et al., 2021), l'ampleur de chaque taille d'effet indirect estimée dans notre cohorte représentait des proportions plus faibles (moins de 45 %) de l'effet total mais était répartie sur un plus grand nombre de médiateurs. Ce résultat suggère que la médiation via l'ADNm placentaire serait davantage polygénique que ce qui a été rapporté dans les études précédentes.

L'analyse de médiation jointe portant sur l'âge gestationnel et le poids de naissance suggère l'existence de relations causales inverses pour les médiateurs AMR situés dans les gènes *COASY* et *BLCAP*, qui peuvent médier une partie de l'effet du tabagisme sur le poids de naissance via un effet inverse sur l'âge gestationnel. Pour clarifier le message

nous avons réalisé un schéma bilan des relations entre le tabagisme, l'âge gestationnel, le poids de naissance et les AMRs médiateurs (figure 3.25).

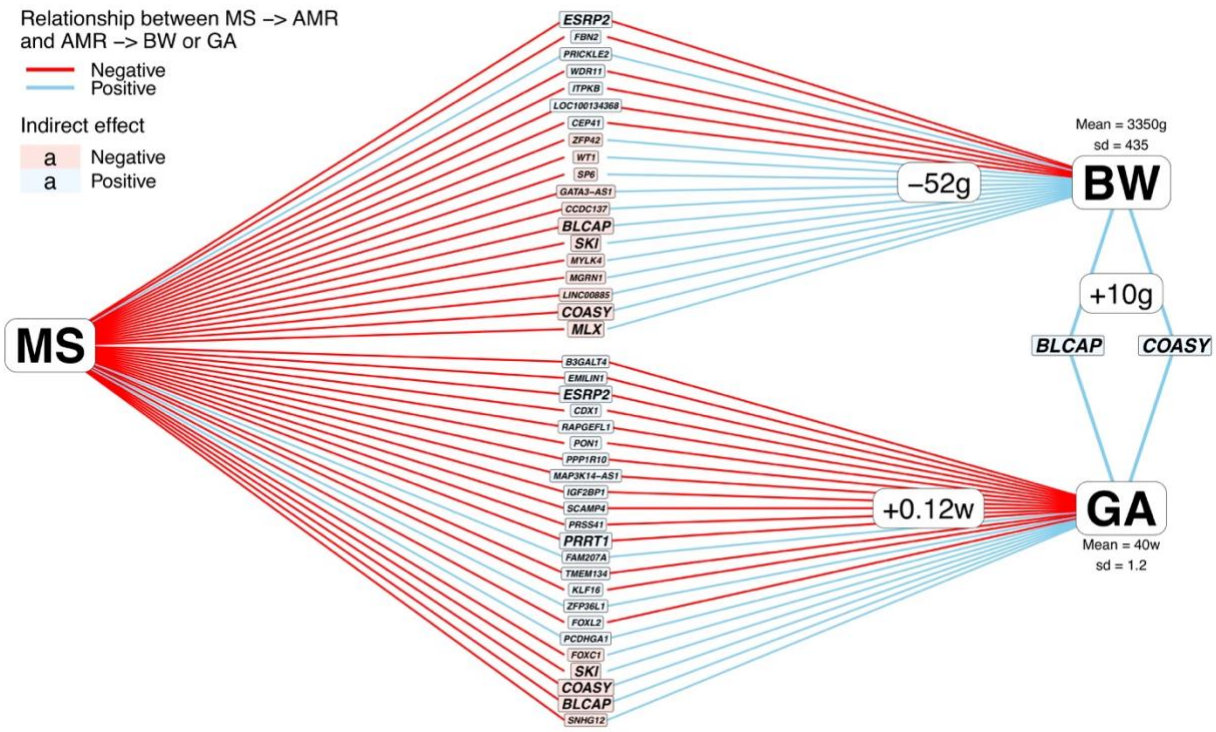

Figure 3.25 : Schéma bilan de l'analyse de médiation des AMRs médiant la relation entre le tabagisme, l'âge gestationnel et le poids à la naissance. *MS = tabagisme maternel, BW = poids à la naissance et GA = âge gestationnel.*

HDMAX2 a permis d'identifier 32 CpGs médiant la relation entre le tabagisme et le poids de naissance, pour lesquels une majorité (20/32) avaient des effets négatifs. Les résultats ont fourni des preuves d'un enrichissement des régions enhancers et d'un appauvrissement en régions promotrices (Rousseaux et al., 2020). Par contre HDMAX2 n'a pas permis de détecter les CpGs identifiés dans les études précédentes (Morales et Cardenas). Plusieurs explications peuvent être avancées. Morales et al n'utilisent pas

dans leur EWAS de méthode d'estimation de la composition cellulaire (ou de méthode à facteurs latents). De plus, pour réaliser leur test de médiation, ils utilisent le test de Sobel qui n'est pas adapté à la grande dimension (Blum et al., 2020). En ce qui concerne Cardenas et al, ils utilisent une puce Illumina 850K tandis que nous utilisons une puce 450K et ils utilisent Refactor pour estimer la composition cellulaire dans leur EWAS. Nous avons constaté via nos études de simulations que cette méthode obtient de faibles performances statistiques.

Au niveau des résultats biologiques, on a pu remarquer que 2 médiateurs ayant des effets indirects négatifs, cg07156115 sur *AHR* et cg20482145 sur *MYH10*, étaient localisés sur des gènes surexprimés dans le placenta par rapport aux autres tissus selon le critère de Chauvenet. De plus, nous avons trouvé 5 CpGs situés sur des gènes au développement selon Gene Ontology : *FGFR2*, *AHR*, *MEIS2*, *SEMA5B* et *CCK*. Les gènes *FGFR2* et *SEMA5B* sont liés au développement d'organismes multicellulaires et à la croissance des organes de développement post-embryonnaires, *AHR* est lié au développement des vaisseaux sanguins, *MEIS2* est lié au développement du cerveau, des yeux et du pancréas, et *CCK* est lié à la migration des neurones.

Côté AMR, *COASY* est lié à de nombreuses voies métaboliques de synthèse et de dégradation, de plus il est co-exprimé avec *MLX*, un AMR proche de 3 Kb de *COASY*. Comme pour les CpGs, nous retrouvons des AMRs situés sur des gènes surexprimés dans le placenta : *FBN2*, *ZFP42* et *KISS1*. 5 AMRs sont liés au développement ou à la croissance des tissus : *FBN2* est lié au développement des yeux, *ZFP42* au

développement des gonades, *KISS1* à la régulation de la sécrétion d'hormone de croissance, *ESRP2* à la voie de signalisation du récepteur du facteur de croissance des fibroblastes et *SKI* au développement de la bouche, du bulbe olfactif, des yeux et des fibres musculaires squelettiques. A noter que les AMRs situés sur *FBN2* et *ESRP2* sont associés à une augmentation du poids à la naissance.

Pour finir, nous avons trouvé un grand nombre de gênes liés à la pré-éclampsie : *NECTIN1* (Ito et al., 2018), *AHR* (Wang et al., 2011), *FGFR2* (Marwa et al., 2016), *COASY* (Martin et al., 2015), *BLCAP* (Li et al., 2020), *SKI* (Martin et al., 2015), *AJAP1* (Yeung et al., 2016) et *SH3BP5* (Kaartokallio et al., 2015). La pré-éclampsie est une maladie de la grossesse qui se traduit par de l'hypertension artérielle. Cette maladie est liée à un tiers des naissances très prématurées et donc la surreprésentation des gènes liés à la pré-éclampsie soutient un effet indirect des médiateurs. A noter qu'une précédente étude de la cohorte EDEN (Abraham et al., 2018), qui portait sur l'impact des polluants atmosphériques sur la méthylation placentaire, avait déjà mis en évidence des gènes liés à la pré-éclampsie (*ADORA2B)* et aux phénomènes d'hypoxie. De plus, le gène *BLCAP* a déjà été identifié comme étant lié au tabagisme maternel dans une EWAS effectuée sur la cohorte EDEN (Rousseaux et al., 2020).

3.5. Conclusion

HDMAX2 est une nouvelle méthode rapide et efficace de mediation de haute dimension. Son application a permis de mieux caractériser la relation entre le tabagisme maternel, la méthylation placentaire, l'âge gestationnel et le poids à la naissance. L'analyse a dépeint une relation causale complexe entre le tabagisme maternel et les issues de grossesse s'appliquant à un plus grand nombre de régions épigénomiques que celles identifiées auparavant, suggérant une architecture polygénique de ces relations. En plus de l'analyse par marqueurs, HDMAX2 a permis l'identification de régions médiatrices (AMR). Les AMRs fournissent des preuves plus robustes que les marqueurs épigénétiques uniques et ont permis la caractérisation de régions médiant les effets du tabagisme maternel à la fois sur l'âge gestationnel et sur le poids de naissance. Trois régions médiatrices communes situées sur les gènes *COASY* et *BLCAP* ont également médié la relation entre l'âge gestationnel et poids de naissance, ce qui suggère une causalité inverse dans la relation entre l'âge gestationnel et le méthylome. En conclusion, notre étude a révélé une complexité insoupçonnée des relations entre le tabagisme maternel durant la grossesse et le poids à la naissance.

4. Conclusion Générale

Les objectifs de la thèse étaient multiples. En premier lieu, nous voulions comprendre les interactions entre une variable d'intérêt, tel qu'une exposition ou un phénotype, et un ensemble de données de séquençage haut débit. Dans un second temps, nous nous sommes attelés à identifier les relations sous-jacentes entre l'exposition au tabagisme prénatal et la santé future des enfants.

Dans le premier chapitre, nous avons présenté une nouvelle méthode, sparse LFMM, permettant de réaliser des études d'associations que ça soit à l'échelle du génome ou de l'épigénome. Cette méthode permet de répondre à de nombreuses problématiques liées aux études d'associations. Elle repose sur des modèles à facteurs ce qui lui permet de prendre en compte la structure des données ainsi mais aussi de sélectionner des marqueurs d'intérêt sans utiliser de tests statistiques. Cette méthode a été validé sur des données simulé et sur des données réelles. L'application sur des données réelles génétiques et épigénétiques a mis en évidence d'anciennes découvertes connues mais aussi a permis de détecter des nouvelles associations biologiquement pertinentes. Ce chapitre a donné lieu à une publication et la méthode est disponible dans le package R *lfmm* téléchargeable sur le CRAN. Pour finir, ce chapitre nous a permis de nous familiariser avec les données de méthylations d'ADN et avec les études d'associations à l'échelle de l'épigénome.

Dans le second chapitre, nous nous attaqué au principal objectif de la thèse. L'état de l'art a mis en évidence un vide méthodologique et une absence de consensus autour des méthodes existantes. Nous nous sommes donc mis comme objectif de développer notre propre méthode de médiation en haute dimension, HDMAX2. Cette méthode repose sur le package LFMM ainsi qu'un nouveau test de médiation rapide et fiable. HDMAX2 a été validé sur des données simulées par rapport à un ensemble de nouvelle méthode. Après le développement de HDMAX2, nous avons pu l'appliquer à notre question biologique qui est l'étude de l'exposition du tabagisme maternel sur la santé future de l'enfant via la méthylation de l'ADN. Nous avons pu mettre en évidence un certain nombre de marqueurs ainsi que des régions médiant les relations suivantes : tabagisme maternel prénatal – méthylation de l'ADN – poids à la naissance et tabagisme maternel prénatal – méthylation de l'ADN – âge gestationnel. HDMAX2 a permis d'estimer les effets indirects totaux ainsi pour l'analyse du poids de naissance nous avons pu constater que 31% des effets passés par la méthylation de l'ADN. Il est apparu que des marqueurs situés sur les gènes *COASY* et *BLCAP* étaient communs aux deux analyses. Ces deux gènes se sont révélés être des médiateurs de l'association âge gestationnel – poids de naissance ce qui laisse penser une relation complexe entre le tabagisme maternel prénatal, la méthylation de l'ADN placentaire, l'âge gestationnel et le poids de naissance. L'application et HDMAX2 ont fait l'objet d'une publication en cours de soumission et notre méthode est disponible dans le dépôt GitHub suivant : *hdmax2*.

Compte tenu de ces résultats, les objectifs de la thèse sont atteints mais on peut tout de même quelques améliorations pouvant être effectuer dans le futur.

Les estimations des effets indirects globaux par le bootstrap ont montré des écarts types très forts. Il serait donc intéressant de continuer à développer cette approche en introduisant un nouveau test statistique permettant de tester ces effets indirects totaux.

Dans la même veine, les effets indirects de chaque médiateur ont été estimé de manière unidimensionnelle et donc sans prendre en compte les autres potentiels médiateurs. Il serait donc intéressant d'utiliser des approches de multi-médiateurs pour estimer les effets indirects des médiateurs potentiels. On pourrait aussi imaginer une approche empirique qui consisterait à inclure en covariable, les premières composantes principales d'une ACP réalisée sur les médiateurs potentiels en excluant le marqueur testé.

HDMAX2 a permis de détecter un certain de marqueur et nous nous sommes principalement intéressés aux gènes les plus proches de ces marqueurs. Mais il serait intéressant d'aller plus loin dans les analyses bio-informatiques. Il faudrait regarder plus précisément le rôle des marqueurs situés sur les régions enhancers et promotrices. De même, nous n'interprétons pas le fait que les marqueurs peuvent hyper ou hypo-méthylés, alors qu'il est connu l'expression génique est impactée de façon différente selon si c'est hyperméthylé ou hypométhylé.

5. Références

Abraham, E., Rousseaux, S., Agier, L., Giorgis-Allemand, L., Tost, J., Galineau, J., Hulin, A., Siroux, V., Vaiman, D., Charles, M.-A., Heude, B., Forhan, A., Schwartz, J., Chuffart, F., Bourova-Flin, E., Khochbin, S., Slama, R., Lepeule, J., 2018. Pregnancy exposure to atmospheric pollution and meteorological conditions and placental DNA methylation. Environ. Int. 118, 334–347. https://doi.org/10.1016/j.envint.2018.05.007

Agha, G., Hajj, H., Rifas-Shiman, S.L., Just, A.C., Hivert, M.-F., Burris, H.H., Lin, X., Litonjua, A.A., Oken, E., DeMeo, D.L., Gillman, M.W., Baccarelli, A.A., 2016. Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. Clin. Epigenetics 8, 118. https://doi.org/10.1186/s13148-016-0285-3

Ananth, C.V., Savitz, D.A., Luther, E.R., 1996. Maternal cigarette smoking as a risk factor for placental abruption, placenta previa, and uterine bleeding in pregnancy. Am. J. Epidemiol. 144, 881–889. https://doi.org/10.1093/oxfordjournals.aje.a009022

Appleton, A.A., Armstrong, D.A., Lesseur, C., Lee, J., Padbury, J.F., Lester, B.M., Marsit, C.J., 2013. Patterning in Placental 11-B Hydroxysteroid Dehydrogenase Methylation According to Prenatal Socioeconomic Adversity. PLOS ONE 8, e74691. https://doi.org/10.1371/journal.pone.0074691

Armitage, P., 1955. Tests for Linear Trends in Proportions and Frequencies. Biometrics 11, 375–386. https://doi.org/10.2307/3001775

Aroian, L.A., Taneja, V.S., Cornwell, L.W., 1978. Mathematical forms of the distribution of the product of two normal variables: Mathematical forms of the distribution. Commun. Stat. - Theory Methods 7, 165–172. https://doi.org/10.1080/03610927808827610

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., Jiang, R., Muliyati, N.W., Zhang, X., Amer, M.A., Baxter, I., Brachi, B., Chory, J., Dean, C., Debieu, M., de Meaux, J., Ecker, J.R., Faure, N., Kniskern, J.M., Jones, J.D.G., Michael, T., Nemri, A., Roux, F., Salt, D.E., Tang, C., Todesco, M., Traw, M.B., Weigel, D., Marjoram, P., Borevitz, J.O., Bergelson, J., Nordborg, M., 2010. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465, 627–631. https://doi.org/10.1038/nature08800

Balding, D.J., 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7, 781–791. https://doi.org/10.1038/nrg1916

Barker, D.J.P., 2007. The origins of the developmental origins theory. J. Intern. Med. 261, 412–417. https://doi.org/10.1111/j.1365-2796.2007.01809.x

Baron, R.M., Kenny, D.A., 1986. The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations 10.

Battram, T., Yousefi, P., Crawford, G., Prince, C., Babei, M.S., Sharp, G., Hatcher, C., Vega-Salas, M.J., Khodabakhsh, S., Whitehurst, O., Langdon, R., Mahoney, L., Elliott, H.R., Mancano, G., Lee, M., Watkins, S.H., Lay, A.C., Hemani, G., Gaunt, T.R., Relton, C.L., Staley, J.R., Suderman, M., 2021. The EWAS Catalog: a database of epigenome-wide association studies. https://doi.org/10.31219/osf.io/837wn

Benjamini, Y., Hochberg, Y., 2000. On the Adaptive Control of the False Discovery Rate in Multiple Testing With Independent Statistics. J. Educ. Behav. Stat. 25, 60–83. https://doi.org/10.3102/10769986025001060

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.

Berlin, I., 2019. Risques associés au tabagisme maternel pendant la grossesse. Attention particulière aux conséquences postanatales. Bull. Académie Natl. Médecine 203, 528–534. https://doi.org/10.1016/j.banm.2019.07.011

Blondel, B., Coulm, B., Bonnet, C., Goffinet, F., Le Ray, C., 2017. Trends in perinatal health in metropolitan France from 1995 to 2016: Results from the French National Perinatal Surveys. J. Gynecol. Obstet. Hum. Reprod. 46, 701–713. https://doi.org/10.1016/j.jogoh.2017.09.002

Blum, M.G.B., Valeri, L., François, O., Cadiou, S., Siroux, V., Lepeule, J., Slama, R., 2020. Challenges Raised by Mediation Analysis in a High-Dimension Setting. Environ. Health Perspect. 128, 055001. https://doi.org/10.1289/EHP6240

Breton, C.V., Siegmund, K.D., Joubert, B.R., Wang, X., Qui, W., Carey, V., Nystad, W., Håberg, S.E., Ober, C., Nicolae, D., Barnes, K.C., Martinez, F., Liu, A., Lemanske, R., Strunk, R., Weiss, S., London, S., Gilliland, F., Raby, B., 2014. Prenatal tobacco smoke exposure is associated with childhood DNA CpG methylation. PloS One 9, e99716. https://doi.org/10.1371/journal.pone.0099716

Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P.L., Amode, R., Guillen, J.A., Riat, H.S., Trevanion, S.J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L.A., Cunningham, F.,

Parkinson, H., 2019. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012. https://doi.org/10.1093/nar/gky1120

Burke, H., Leonardi-Bee, J., Hashim, A., Pine-Abata, H., Chen, Y., Cook, D.G., Britton, J.R., McKeever, T.M., 2012. Prenatal and passive smoke exposure and incidence of asthma and wheeze: systematic review and meta-analysis. Pediatrics 129, 735–744. https://doi.org/10.1542/peds.2011-2196

Byzova, M.V., Franken, J., Aarts, M.G.M., de Almeida-Engler, J., Engler, G., Mariani, C., Van Lookeren Campagne, M.M., Angenent, G.C., 1999. Arabidopsis STERILE APETALA, a multifunctional gene regulating inflorescence, flower, and ovule development. Genes Dev. 13, 1002–1014. https://doi.org/10.1101/gad.13.8.1002

Cai, J.-F., Candes, E.J., Shen, Z., 2008. A Singular Value Thresholding Algorithm for Matrix Completion. ArXiv08103286 Math.

Cardenas, A., Lutz, S.M., Everson, T.M., Perron, P., Bouchard, L., Hivert, M.-F., 2019. Placental DNA Methylation Mediates the Association of Prenatal Maternal Smoking on Birth Weight. Am. J. Epidemiol. https://doi.org/10.1093/aje/kwz184

Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M., 2008. High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. J. Am. Stat. Assoc. 103, 1438–1456. https://doi.org/10.1198/016214508000000869

Caye, K., Jumentier, B., Lepeule, J., François, O., 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Mol. Biol. Evol. 36, 852–860. https://doi.org/10.1093/molbev/msz008

Chu, 2016. Maternal smoking during pregnancy and risk of childhood neuroblastoma: Systematic review and meta-analysis [WWW Document]. URL https://www.cancerjournal.net/article.asp?issn=0973-1482;year=2016;volume=12;issue=2;spage=999;epage=1005;aulast=Chu (accessed 2.2.22).

Clifford, A., Lang, L., Chen, R., 2012. Effects of maternal cigarette smoking during pregnancy on cognitive parameters of children and young adults: a literature review. Neurotoxicol. Teratol. 34, 560–570. https://doi.org/10.1016/j.ntt.2012.09.004

Cnattingius, S., 2004. The epidemiology of smoking during pregnancy: smoking prevalence, maternal

characteristics, and pregnancy outcomes. Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob. 6 Suppl 2, S125-140. https://doi.org/10.1080/14622200410001669187

Cupul-Uicab, L.A., Skjaerven, R., Haug, K., Melve, K.K., Engel, S.M., Longnecker, M.P., 2012. In Utero Exposure to Maternal Tobacco Smoke and Subsequent Obesity, Hypertension, and Gestational Diabetes Among Women in The MoBa Cohort. Environ. Health Perspect. 120, 355–360. https://doi.org/10.1289/ehp.1103789

Dai, J.Y., Stanford, J.L., LeBlanc, M., 2020. A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses. J. Am. Stat. Assoc. 1–16. https://doi.org/10.1080/01621459.2020.1765785

Decamps, C., Privé, F., Bacher, R., Jost, D., Waguet, A., Achard, S., Amblard, E., Bacher, R., Bergmann, F., Blum, M., Blum, Y., Bottaz-Bosson, G., Broseus, L., Chuffart, F., Decamps, C., Devijver, E., Durif, G., Feofanov, V., Houseman, E.A., Gallopin, M., Jedynak, P., Jonchere, V., van de Geer, E., Jumentier, B., Kaoma, T., Lurie, E., Lutsik, P., Markowski, J., Melnykova, A., Merlevede, J., Nazarov, P., Nguyen, N.H., Permiakova, O., Privé, F., Richard, M., Rolland, M., Scherer, M., Spill, Y., Houseman, E.A., Lurie, E., Lutsik, P., Milosavljevic, A., Scherer, M., Blum, M.G.B., Richard, M., HADACA consortium, 2020. Guidelines for cell-type heterogeneity quantification based on a comparative analysis of reference-free DNA methylation deconvolution software. BMC Bioinformatics 21, 16. https://doi.org/10.1186/s12859-019-3307-2

Demiguel, V., 2018. ÉVOLUTION DE LA CONSOMMATION DE TABAC À L'OCCASION D'UNE GROSSESSE EN FRANCE EN 2016 / EVOLUTION OF TOBACCO USE DURING PREGNANCY IN FRANCE IN 2016 10.

Devlin, B., Roeder, K., 1999. Genomic control for association studies. Biometrics 55, 997–1004. https://doi.org/10.1111/j.0006-341x.1999.00997.x

Djordjilović, V., Hemerik, J., Thoresen, M., 2020. On optimal two-stage testing of multiple mediators. ArXiv200702844 Stat.

Djordjilović, V., Page, C.M., Gran, J.M., Nøst, T.H., Sandanger, T.M., Veierød, M.B., Thoresen, M., 2019. Global test for high-dimensional mediation: Testing groups of potential mediators. Stat. Med. https://doi.org/10.1002/sim.8199

Eckart, C., Young, G., 1936. The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218. https://doi.org/10.1007/BF02288367

Efron, B., 2004. Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. J. Am. Stat. Assoc. 99, 96–104.

Everson, T.M., Vives-Usano, M., Seyve, E., Cardenas, A., Lacasaña, M., Craig, J.M., Lesseur, C., Baker, E.R., Fernandez-Jimenez, N., Heude, B., Perron, P., Gónzalez-Alzaga, B., Halliday, J., Deyssenroth, M.A., Karagas, M.R., Íñiguez, C., Bouchard, L., Carmona-Sáez, P., Loke, Y.J., Hao, K., Belmonte, T., Charles, M.A., Martorell-Marugán, J., Muggli, E., Chen, J., Fernández, M.F., Tost, J., Gómez-Martín, A., London, S.J., Sunyer, J., Marsit, C.J., Lepeule, J., Hivert, M.-F., Bustamante, M., 2019. Placental DNA methylation signatures of maternal smoking during pregnancy and potential impacts on fetal growth (preprint). Genomics. https://doi.org/10.1101/663567

Fan, J., Lv, J., 2008. Sure independence screening for ultrahigh dimensional feature space. J. R. Stat. Soc. Ser. B Stat. Methodol. 70, 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x

Feinberg, A.P., 2018. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. N. Engl. J. Med. 378, 1323–1334. https://doi.org/10.1056/NEJMra1402513

Fogelman, K., Manor, O., 1988. Smoking in pregnancy and development into early adulthood. https://doi.org/10.1136/bmj.297.6658.1233

François, O., Caye, K., 2018. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. Mol. Ecol. Resour. 18, 789–797. https://doi.org/10.1111/1755-0998.12892

Frichot, E., Schoville, S.D., Bouchard, G., François, O., 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Mol. Biol. Evol. 30, 1687–1699. https://doi.org/10.1093/molbev/mst063

Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., 2007. Pathwise coordinate optimization. Ann. Appl. Stat. 1, 302–332. https://doi.org/10.1214/07-AOAS131

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization Paths for Generalized Linear Models via Coordinate Descent. J. Stat. Softw. 33. https://doi.org/10.18637/jss.v033.i01

Gao, Y., Yang, H., Fang, R., Zhang, Y., Goode, E.L., Cui, Y., 2019. Testing Mediation Effects in High-Dimensional Epigenetic Studies. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.01195

Gluckman, P.D., Hanson, M.A., Buklijas, T., 2010. A conceptual framework for the developmental origins

of health and disease. J. Dev. Orig. Health Dis. 1, 6–18. https://doi.org/10.1017/S2040174409990171

Goodman, L.A., 1960. On the Exact Variance of Products. J. Am. Stat. Assoc. 55, 708. https://doi.org/10.2307/2281592

Hanzawa, Y., Takahashi, T., Komeda, Y., 1997. ACL5: an Arabidopsis gene required for internodal elongation after flowering. Plant J. 12, 863–874. https://doi.org/10.1046/j.1365-313X.1997.12040863.x

Hayes, A.F., 2009. Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. Commun. Monogr. 76, 408–420. https://doi.org/10.1080/03637750903310360

Heude, B., Forhan, A., Slama, R., Douhaud, L., Bedel, S., Saurel-Cubizolles, M.-J., Hankard, R., Thiebaugeorges, O., De Agostini, M., Annesi-Maesano, I., Kaminski, M., Charles, M.-A., 2016. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. Int. J. Epidemiol. 45, 353–363. https://doi.org/10.1093/ije/dyv151

Hoggart, C.J., Clark, T.G., De Iorio, M., Whittaker, J.C., Balding, D.J., 2008. Genome-wide significance for dense SNP and resequencing data. Genet. Epidemiol. 32, 179–185. https://doi.org/10.1002/gepi.20292

Houseman, E.A., Kile, M.L., Christiani, D.C., Ince, T.A., Kelsey, K.T., Marsit, C.J., 2016. Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. BMC Bioinformatics 17. https://doi.org/10.1186/s12859-016-1140-4

Houseman, E.A., Molitor, J., Marsit, C.J., 2014. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics 30, 1431–1439. https://doi.org/10.1093/bioinformatics/btu029

Huang, Y.-T., Pan, W.-C., 2016. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators: Hypothesis Test of Mediation Effect in Causal Mediation Model with High-Dimensional Continuous Mediators. Biometrics 72, 402–413. https://doi.org/10.1111/biom.12421

Imai, K., Keele, L., Tingley, D., 2010. A general approach to causal mediation analysis. Psychol. Methods 15, 309–334. https://doi.org/10.1037/a0020761

Ino, T., 2010. Maternal smoking during pregnancy and offspring obesity: meta-analysis. Pediatr. Int. Off. J. Jpn. Pediatr. Soc. 52, 94–99. https://doi.org/10.1111/j.1442-200X.2009.02883.x

Ito, M., Nishizawa, H., Tsutsumi, M., Kato, A., Sakabe, Y., Noda, Y., Ohwaki, A., Miyazaki, J., Kato, T.,

Shiogama, K., Sekiya, T., Kurahashi, H., Fujii, T., 2018. Potential role for nectin-4 in the pathogenesis of pre-eclampsia: a molecular genetic study. BMC Med. Genet. 19, 166. https://doi.org/10.1186/s12881-018-0681-y

Jedynak, P., Maitre, L., Guxens, M., Gützkow, K.B., Julvez, J., López-Vicente, M., Sunyer, J., Casas, M., Chatzi, L., Gražulevičienė, R., Kampouri, M., McEachan, R., Mon-Williams, M., Tamayo, I., Thomsen, C., Urquiza, J., Vafeiadi, M., Wright, J., Basagaña, X., Vrijheid, M., Philippat, C., 2021. Prenatal exposure to a wide range of environmental chemicals and child behaviour between 3 and 7 years of age – An exposome-based approach in 5 European cohorts. Sci. Total Environ. 763, 144115. https://doi.org/10.1016/j.scitotenv.2020.144115

Joubert, B.R., Håberg, S.E., Nilsen, R.M., Wang, X., Vollset, S.E., Murphy, S.K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L.A., Ueland, P.M., Wu, M.C., Nystad, W., Bell, D.A., Peddada, S.D., London, S.J., 2012. 450K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy. Environ. Health Perspect. 120, 1425–1431. https://doi.org/10.1289/ehp.1205412

Kaartokallio, T., Cervera, A., Kyllönen, A., Laivuori, K., Kere, J., Laivuori, H., FINNPEC Core Investigator Group, 2015. Gene expression profiling of pre-eclamptic placentae by RNA sequencing. Sci. Rep. 5, 14107. https://doi.org/10.1038/srep14107

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M., 2021. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 49, D545–D551. https://doi.org/10.1093/nar/gkaa970

Kapustin, R.V., Drobintseva, A.O., Alekseenkova, E.N., Onopriychuk, A.R., Arzhanova, O.N., Polyakova, V.O., Kvetnoy, I.M., 2020. Placental protein expression of kisspeptin-1 (KISS1) and the kisspeptin-1 receptor (KISS1R) in pregnancy complicated by diabetes mellitus or preeclampsia. Arch. Gynecol. Obstet. 301, 437–445. https://doi.org/10.1007/s00404-019-05408-1

Kaushal, A., Zhang, H., Karmaus, W.J., Wang, J.S., 2015. Which methods to choose to correct cell types in genome-scale blood-derived DNA methylation data? BMC Bioinformatics 16, P7. https://doi.org/10.1186/1471-2105-16-S15-P7

Küpers, L.K., Xu, X., Jankipersadsing, S.A., Vaez, A., la Bastide-van Gemert, S., Scholtens, S., Nolte, I.M., Richmond, R.C., Relton, C.L., Felix, J.F., Duijts, L., van Meurs, J.B., Tiemeier, H., Jaddoe, V.W., Wang, X., Corpeleijn, E., Snieder, H., 2015. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. Int. J. Epidemiol. 44, 1224–1237.

https://doi.org/10.1093/ije/dyv048

Lee, S., Sun, W., Wright, F.A., Zou, F., 2017. An improved and explicit surrogate variable analysis procedure by coefficient adjustment. Biometrika 104, 303–316. https://doi.org/10.1093/biomet/asx018

Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., Storey, J.D., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics 28, 882–883. https://doi.org/10.1093/bioinformatics/bts034

Leek, J.T., Storey, J.D., 2008. A general framework for multiple testing dependence. Proc. Natl. Acad. Sci. 105, 18718–18723. https://doi.org/10.1073/pnas.0808709105

Leek, J.T., Storey, J.D., 2007. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet. 3, 12.

Lermenier-Jeannet, A., 2014. Le tabac en France : un bilan des années 2004-2014.

Li, Yingying, Cui, S., Shi, W., Yang, B., Yuan, Y., Yan, S., Li, Ying, Xu, Y., Zhang, Z., Linlin Zhang, null, 2020. Differential placental methylation in preeclampsia, preterm and term pregnancies. Placenta 93, 56–63. https://doi.org/10.1016/j.placenta.2020.02.009

Lotterhos, K.E., 2019. The Effect of Neutral Recombination Variation on Genome Scans for Selection. G3 Bethesda Md 9, 1851–1867. https://doi.org/10.1534/g3.119.400088

Lutsik, P., Slawski, M., Gasparoni, G., Vedeneev, N., Hein, M., Walter, J., 2017. MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. Genome Biol. 18. https://doi.org/10.1186/s13059-017-1182-6

MacKinnon, D.P., Fairchild, A.J., Fritz, M.S., 2007. Mediation Analysis. Annu. Rev. Psychol. 58, 593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542

MacKinnon, D.P., Lockwood, C.M., Hoffman, J.M., West, S.G., 2002. A Comparison of Methods to Test Mediation and Other Intervening Variable Effects 35.

MacKinnon, D.P., Lockwood, C.M., Williams, J., 2004. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. Multivar. Behav. Res. 39, 99–128. https://doi.org/10.1207/s15327906mbr3901_4

Martin, E., Ray, P.D., Smeester, L., Grace, M.R., Boggess, K., Fry, R.C., 2015. Epigenetics and Preeclampsia: Defining Functional Epimutations in the Preeclamptic Placenta Related to the TGF-β Pathway. PLOS ONE 10, e0141294. https://doi.org/10.1371/journal.pone.0141294

Marwa, B.A.G., Raguema, N., Zitouni, H., Feten, H.B.A., Olfa, K., Elfeleh, R., Almawi, W., Mahjoub, T., 2016. FGF1 and FGF2 mutations in preeclampsia and related features. Placenta 43, 81–85. https://doi.org/10.1016/j.placenta.2016.05.007

Mishra, B., Meyer, G., Bach, F., Sepulchre, R., 2013. Low-rank optimization with trace norm penalty. ArXiv11122318 Cs Math.

Morales, E., Vilahur, N., Salas, L.A., Motta, V., Fernandez, M.F., Murcia, M., Llop, S., Tardon, A., Fernandez-Tardon, G., Santa-Marina, L., Gallastegui, M., Bollati, V., Estivill, X., Olea, N., Sunyer, J., Bustamante, M., 2016. Genome-wide DNA methylation study in human placenta identifies novel loci associated with maternal smoking during pregnancy. Int. J. Epidemiol. 45, 1644–1655. https://doi.org/10.1093/ije/dyw196

Murphy, V.E., Smith, R., Giles, W.B., Clifton, V.L., 2006. Endocrine Regulation of Human Fetal Growth: The Role of the Mother, Placenta, and Fetus. Endocr. Rev. 27, 141–169. https://doi.org/10.1210/er.2005-0011

Nabet, C., Ancel, P.-Y., Burguet, A., Kaminski, M., 2005. Smoking during pregnancy and preterm birth according to obstetric history: French national perinatal surveys. Paediatr. Perinat. Epidemiol. 19, 88–96. https://doi.org/10.1111/j.1365-3016.2005.00639.x

Nakamura, A., François, O., Lepeule, J., 2021. Epigenetic Alterations of Maternal Tobacco Smoking during Pregnancy: A Narrative Review. Int. J. Environ. Res. Public. Health 18, 5083. https://doi.org/10.3390/ijerph18105083

Nishimura, N., Tsuchiya, W., Moresco, J.J., Hayashi, Y., Satoh, K., Kaiwa, N., Irisa, T., Kinoshita, T., Schroeder, J.I., Yates, J.R., Hirayama, T., Yamazaki, T., 2018. Control of seed dormancy and germination by DOG1-AHG1 PP2C phosphatase complex via binding to heme. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-04437-9

Onuchic, V., Hartmaier, R.J., Boone, D.N., Samuels, M.L., Patel, R.Y., White, W.M., Garovic, V.D., Oesterreich, S., Roth, M.E., Lee, A.V., Milosavljevic, A., 2016. Epigenomic Deconvolution of Breast Tumors

Reveals Metabolic Coupling between Constituent Cell Types. Cell Rep. 17, 2075–2086. https://doi.org/10.1016/j.celrep.2016.10.057

Papatheodorou, I., Moreno, P., Manning, J., Fuentes, A.M.-P., George, N., Fexova, S., Fonseca, N.A., Füllgrabe, A., Green, M., Huang, N., Huerta, L., Iqbal, H., Jianu, M., Mohammed, S., Zhao, L., Jarnuczak, A.F., Jupp, S., Marioni, J., Meyer, K., Petryszak, R., Prada Medina, C.A., Talavera-López, C., Teichmann, S., Vizcaino, J.A., Brazma, A., 2020. Expression Atlas update: from tissues to single cells. Nucleic Acids Res. 48, D77–D83. https://doi.org/10.1093/nar/gkz947

Patterson, N., Price, A.L., Reich, D., 2006. Population Structure and Eigenanalysis. PLOS Genet. 2, e190. https://doi.org/10.1371/journal.pgen.0020190

Peiffer, G., Underner, M., Perriot, J., 2018. Les effets respiratoires du tabagisme. Rev. Pneumol. Clin., Numéro spécial “Tabacologie” 74, 133–144. https://doi.org/10.1016/j.pneumo.2018.04.009

Preacher, K.J., Hayes, A.F., 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behav. Res. Methods Instrum. Comput. 36, 717–731. https://doi.org/10.3758/BF03206553

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D., 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38, 904–909. https://doi.org/10.1038/ng1847

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am. J. Hum. Genet. 81, 559–575.

Rahmani, E., Zaitlen, N., Baran, Y., Eng, C., Hu, D., Galanter, J., Oh, S., Burchard, E.G., Eskin, E., Zou, J., Halperin, E., 2016. Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. Nat. Methods 13, 443–445. https://doi.org/10.1038/nmeth.3809

Rakyan, V.K., Down, T.A., Balding, D.J., Beck, S., 2011. Epigenome-wide association studies for common human diseases. Nat. Rev. Genet. 12, 529–541. https://doi.org/10.1038/nrg3000

Rich-Edwards, J.W., Stampfer, M.J., Manson, J.E., Rosner, B., Hankinson, S.E., Colditz, G.A., Hennekens, C.H., Willet, W.C., 1997. Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. BMJ 315, 396–400. https://doi.org/10.1136/bmj.315.7105.396

Rousseaux, S., Seyve, E., Chuffart, F., Bourova-Flin, E., Benmerad, M., Charles, M.-A., Forhan, A., Heude, B., Siroux, V., Slama, R., Tost, J., Vaiman, D., Khochbin, S., Lepeule, J., the EDEN mother-child cohort study group, 2019. Maternal exposure to cigarette smoking induces immediate and durable changes in placental DNA methylation affecting enhancer and imprinting control regions (preprint). Genomics. https://doi.org/10.1101/852186

Rousseaux, S., Seyve, E., Chuffart, F., Bourova-Flin, E., Benmerad, M., Charles, Marie-Aline, Forhan, Anne, Heude, Barbara, Siroux, V., Slama, Remy, Tost, J., Vaiman, D., Khochbin, S., Lepeule, Johanna, Annesi-Maesano, I., Bernard, J.Y., Botton, J., Charles, M-A, Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., de Agostini, M., Foliguet, B., Forhan, A., Fritel, X., Germa, A., Goua, V., Hankard, R., Heude, B., Kaminski, M., Larroque, B., Lelong, N., Lepeule, J., Magnin, G., Marchand, L., Nabet, C., Pierre, F., Slama, R., Saurel-Cubizolles, M.J., Schweitzer, M., Thiebaugeorges, O., the EDEN Mother-Child Cohort Study Group, 2020. Immediate and durable effects of maternal tobacco consumption alter placental DNA methylation in enhancer and imprinted gene-containing regions. BMC Med. 18, 306. https://doi.org/10.1186/s12916-020-01736-1

Rumrich, I.K., Viluksela, M., Vähäkangas, K., Gissler, M., Surcel, H.-M., Hänninen, O., 2016. Maternal Smoking and the Risk of Cancer in Early Life - A Meta-Analysis. PloS One 11, e0165040. https://doi.org/10.1371/journal.pone.0165040

Sado, T., Naruse, K., Noguchi, T., Haruta, S., Yoshida, S., Tanase, Y., Kitanaka, T., Oi, H., Kobayashi, H., 2011. Inflammatory pattern recognition receptors and their ligands: factors contributing to the pathogenesis of preeclampsia. Inflamm. Res. 60, 509–520. https://doi.org/10.1007/s00011-011-0319-4

Sampson, J.N., Boca, S.M., Moore, S.C., Heller, R., 2018. FWER and FDR control when testing multiple mediators. Bioinformatics 34, 2418–2424. https://doi.org/10.1093/bioinformatics/bty064

She, Y., Chen, K., 2017. Robust reduced-rank regression. Biometrika 104, 633–647. https://doi.org/10.1093/biomet/asx032

Sheldon, C.C., Rouse, D.T., Finnegan, E.J., Peacock, W.J., Dennis, E.S., 2000. The molecular basis of vernalization: The central role of FLOWERING LOCUS C (FLC). PLANT Biol. 97, 6.

Shimanuki, Y., Mitomi, H., Fukumura, Y., Makino, S., Itakura, A., Yao, T., Takeda, S., 2015. Alteration of Delta-like ligand 1 and Notch 1 receptor in various placental disorders with special reference to early onset preeclampsia. Hum. Pathol. 46, 1129–1137. https://doi.org/10.1016/j.humpath.2015.03.013

Sobel, M.E., 1982. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociol. Methodol. 13, 290. https://doi.org/10.2307/270723

Song, Y., Zhou, X., Zhang, M., Zhao, W., Liu, Y., Kardia, S.L.R., Roux, A.V.D., Needham, B.L., Smith, J.A., Mukherjee, B., 2020. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. Biometrics 76, 700–710. https://doi.org/10.1111/biom.13189

Storey, J.D., 2002. A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 479–498. https://doi.org/10.1111/1467-9868.00346

Strimmer, K., 2008. A unified approach to false discovery rate estimation. BMC Bioinformatics 9, 303. https://doi.org/10.1186/1471-2105-9-303

Suter, M., Abramovici, A., Showalter, L., Hu, M., Shope, C.D., Varner, M., Aagaard-Tillery, K., 2010. In utero tobacco exposure epigenetically modifies placental CYP1A1 expression. Metabolism. 59, 1481–1490. https://doi.org/10.1016/j.metabol.2010.01.013

Suter, M., Ma, J., Harris, A., Patterson, L., Brown, K.A., Shope, C., Showalter, L., Abramovici, A., Aagaard-Tillery, K.M., 2011. Maternal tobacco use modestly alters correlated epigenome-wide placental DNA methylation and gene expression. Epigenetics 6, 1284–1294. https://doi.org/10.4161/epi.6.11.17819

Teschendorff, A.E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., Beck, S., 2013. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinforma. Oxf. Engl. 29, 189–196. https://doi.org/10.1093/bioinformatics/bts680

Teschendorff, A.E., Zheng, S.C., 2017. Cell-type deconvolution in epigenome-wide association studies: a review and recommendations. Epigenomics 9, 757–768. https://doi.org/10.2217/epi-2016-0153

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. Ser. B Methodol. 58, 267–288.

Tobi, E.W., Slieker, R.C., Luijk, R., Dekkers, K.F., Stein, A.D., Xu, K.M., Biobank-based Integrative Omics Studies Consortium, Slagboom, P.E., van Zwet, E.W., Lumey, L.H., Heijmans, B.T., 2018. DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. Sci. Adv. 4, eaao4364. https://doi.org/10.1126/sciadv.aao4364

Triche, T.J., 2014. FDb.InfiniumMethylation.hg19: Annotation package for Illumina Infinium DNA methylation probes. R package version 2.2.0.

van Kesteren, E.-J., Oberski, D.L., 2019. Exploratory Mediation Analysis with Many Potential Mediators. Struct. Equ. Model. Multidiscip. J. 26, 710–723. https://doi.org/10.1080/10705511.2019.1588124

VanderWeele, T., 2015. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford University Press.

VanderWeele, T.J., 2016. Mediation Analysis: A Practitioner's Guide. Annu. Rev. Public Health 37, 17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402

Wang, J., Zhao, Q., Hastie, T., Owen, A.B., 2017. Confounder adjustment in multiple hypothesis testing. Ann. Stat. 45, 1863–1894. https://doi.org/10.1214/16-AOS1511

Wang, K., Zhou, Q., He, Q., Tong, G., Zhao, Z., Duan, T., 2011. The possible role of AhR in the protective effects of cigarette smoke on preeclampsia. Med. Hypotheses 77, 872–874. https://doi.org/10.1016/j.mehy.2011.07.061

Wang, M.H., Cordell, H.J., Van Steen, K., 2019. Statistical methods for genome-wide association studies. Semin. Cancer Biol. 55, 53–60. https://doi.org/10.1016/j.semcancer.2018.04.008

Wu, T.T., Chen, Y.F., Hastie, T., Sobel, E., Lange, K., 2009. Genome-wide association analysis by lasso penalized logistic regression. Bioinforma. Oxf. Engl. 25, 714–721. https://doi.org/10.1093/bioinformatics/btp041

Xu, R., Hong, X., Zhang, B., Huang, W., Hou, W., Wang, G., Wang, X., Igusa, T., Liang, L., Ji, H., 2021. DNA methylation mediates the effect of maternal smoking on offspring birthweight: a birth cohort study of multi-ethnic US mother–newborn pairs. Clin. Epigenetics 13, 47. https://doi.org/10.1186/s13148-021-01032-6

Xu, Z., Niu, L., Li, L., Taylor, J.A., 2016. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 44, e20. https://doi.org/10.1093/nar/gkv907

Yeung, K.R., Chiu, C.L., Pidsley, R., Makris, A., Hennessy, A., Lind, J.M., 2016. DNA methylation profiles in preeclampsia and healthy control placentas. Am. J. Physiol. Heart Circ. Physiol. 310, H1295-1303. https://doi.org/10.1152/ajpheart.00958.2015

Zeng, P., Shao, Z., Zhou, X., 2021. Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. Comput. Struct. Biotechnol. J. 19, 3209–3224. https://doi.org/10.1016/j.csbj.2021.05.042

Zeng, P., Zhou, X., Huang, S., 2017. Prediction of gene expression with cis-SNPs using mixed models and regularization methods. BMC Genomics 18. https://doi.org/10.1186/s12864-017-3759-6

Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38, 894–942. https://doi.org/10.1214/09-AOS729

Zhang, F., Chen, W., Zhu, Z., Zhang, Q., Nabais, M.F., Qi, T., Deary, I.J., Wray, N.R., Visscher, P.M., McRae, A.F., Yang, J., 2019. OSCA: a tool for omic-data-based complex trait analysis. Genome Biol. 20, 107. https://doi.org/10.1186/s13059-019-1718-z

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., Liu, L., 2016. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics 32, 3150–3154. https://doi.org/10.1093/bioinformatics/btw351

Zhou, X., Carbonetto, P., Stephens, M., 2013. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. PLoS Genet. 9, e1003264. https://doi.org/10.1371/journal.pgen.1003264

Zhou, X., Stephens, M., 2012. Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. 44, 821–824. https://doi.org/10.1038/ng.2310

6. Annexe

LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies

Kevin Caye,[1] Basile Jumentier,[1] Johanna Lepeule,[2] and Olivier François*,[1]

[1]Université Grenoble-Alpes, Centre National de la Recherche Scientifique, Grenoble INP, TIMC-IMAG CNRS UMR 5525, Grenoble 38000, France

[2]Université Grenoble-Alpes, Institut National de la Santé et de la Recherche Médicale, Centre National de la Recherche Scientifique, Institute for Advanced Biosciences, INSERM U 1209 - CNRS UMR 5309, Grenoble 38000, France

*Corresponding author: E-mail: olivier.francois@grenoble-inp.fr.

Associate editor: Joanna Kelley

Abstract

Gene-environment association (GEA) studies are essential to understand the past and ongoing adaptations of organisms to their environment, but those studies are complicated by confounding due to unobserved demographic factors. Although the confounding problem has recently received considerable attention, the proposed approaches do not scale with the high-dimensionality of genomic data. Here, we present a new estimation method for latent factor mixed models (LFMMs) implemented in an upgraded version of the corresponding computer program. We developed a least-squares estimation approach for confounder estimation that provides a unique framework for several categories of genomic data, not restricted to genotypes. The speed of the new algorithm is several order faster than existing GEA approaches and then our previous version of the LFMM program. In addition, the new method outperforms other fast approaches based on principal component or surrogate variable analysis. We illustrate the program use with analyses of the 1000 Genomes Project data set, leading to new findings on adaptation of humans to their environment, and with analyses of DNA methylation profiles providing insights on how tobacco consumption could affect DNA methylation in patients with rheumatoid arthritis.

Software availability: Software is available in the R package lfmm at https://bcm-uga.github.io/lfmm/.

Key words: **gene-environment association, local adaptation, ecological genomics, confounding factors, statistical methods.**

Article

Introduction

Association studies have been extensively used to identify genes or molecular markers associated with disease states, exposure levels or phenotypic traits. Given a large number of molecular markers, the objective of those studies is to test whether any of the markers exhibits significant correlation with a primary variable of interest. Among those methods, gene-environment association (GEA) studies propose to test for correlation with ecological gradients in order to detect genomic signatures of local adaptation (Savolainen et al. 2013).

Although they bring useful information on the molecular targets of selection, GEA studies suffer from the problem of confounding. This problem arises when there exist unobserved variables that correlate both with the primary variables and genomic data (Wang et al. 2017). Recently, several model-based approaches have been introduced to evaluate GEA while correcting for unobserved demographic processes and population structure. Those methods include the programs BAYENV (Günther and Coop 2013), BAYPASS (Gautier 2015), BAYESCENV (Villemereuil and Gaggiotti 2015), and latent factor mixed model (LFMM) (Frichot et al. 2013; Frichot and François 2015). The use of those methods has become popular in ecological genomics, and several surveys have shown that they are robust to departure from their model assumptions (De Mita et al. 2013; Villemereuil et al. 2014; Lotterhos and Whitlock 2015; Rellstab et al. 2015). One drawback of those approaches, however, is to rely on Markov chain Monte Carlo algorithms or Bayesian bootstrap methods to perform parameter inference and statistical testing. Monte Carlo methods are flexible and allow complex models to be implemented in a computer program, but they can be highly intensive and they run slowly. Although some programs have parallel versions for multiprocessor systems, there is a need to develop fast and accurate methods that scale with the very large dimensions of genomic data sets and save computer energy.

In this study, we present a new version of the LFMM algorithm based on the solution of a regularized least-squares minimization problem. In addition, the new models are extended to handle data other than genotypes, and to perform multivariate regressions with more than one explanatory variable or a more general design matrix. Until now, GEAs have mainly focused on single-nucleotide polymorphisms (SNPs) by examining genetic variants in different individuals. In recent years, other categories of data have emerged and become of specific interest. For example, some

epigenome-wide association studies (EWAS) measure DNA methylation levels in different individuals to derive associations between epigenetic variation and exposure levels or phenotypes (Rakyan et al. 2011; Teschendorff and Relton 2018). Here, we extend the definition of the LFMM data matrix to DNA methylation profiles and other molecular markers within a unified framework (Leek and Storey 2007; Carvalho et al. 2008). We present our new LFMM method in the next section. Then we demonstrate that our new method is several orders faster than its previous Bayesian version without loss of power or precision. In a GEA study of individuals from the 1000 Genomes Project, the new program detects genes linked to climate in humans. In an EWAS of patients with rheumatoid arthritis (RA), it identifies a set of genes for which DNA methylation potentially mediates the effect of tobacco consumption on the disease phenotype.

New Approach

GEA methods evaluate associations between the elements of a response matrix, **Y**, and some variables of interest, called "environmental" or "primary" variables, **X**, measured for n individuals. The response matrix records data for the n individuals, which often correspond to genotypes measured from p genetic markers. Here we extend the definition of **Y** to DNA methylation profiles (beta-normalized values) or gene-expression data. Nuisance variables such as observed confounders can be included in the **X** matrix, which dimension is then $n \times d$, where d represents the total number of primary and nuisance variables.

LFMMs are regression models combining fixed and latent effects as follows

$$\mathbf{Y} = \mathbf{X}\mathbf{B}^T + \mathbf{W} + \mathbf{E}. \quad (1)$$

The fixed effect sizes are recorded in the **B** matrix, which has dimension $p \times d$. The **E** matrix represents residual errors, and it has the same dimensions as the response matrix. The matrix **W** is a "latent matrix" of rank K, defined by K latent factors where K can be determined by model choice procedures (Leek and Storey 2007; Frichot et al. 2013). The K latent factors represent unobserved confounders which are modeled through an $n \times K$ matrix, **U**. The matrix **U** is obtained from a singular value decomposition (SVD) of the matrix **W** as follows

$$\mathbf{W} = \mathbf{U}\mathbf{V}^T,$$

where **V** is a $p \times K$ matrix of loadings (Eckart and Young 1936). The **U** and **V** matrices are unique up to arbitrary signs for the factors and loadings.

L^2-Regularized Least-Squares Problem

In our new version of LFMM, the statistical estimates of latent factors and environmental effects are based on least-squares minimization. More specifically, statistical estimates of the parameter matrices **U**, **V**, **B** are computed after minimizing the following penalized loss function

$$\mathcal{L}_{\text{ridge}}(\mathbf{U},\mathbf{V},\mathbf{B}) = ||\mathbf{Y} - \mathbf{U}\mathbf{V}^T - \mathbf{X}\mathbf{B}^T||_F^2 + \lambda||\mathbf{B}||_2^2, \quad \lambda > 0, \quad (2)$$

where $||.||_F$ is the Frobenius norm, $||.||_2$ is the L^2 norm, and λ is a regularization parameter. A positive value of the regularization parameter is necessary for identifying the parameter matrices $\mathbf{W} = \mathbf{U}\mathbf{V}^T$ and **B**. For $\lambda = 0$, the solutions of the least-squares problem are not defined unequivocally and infinitely many solutions exist. A basic algorithm that computes a low rank approximation of the response matrix using their first K principal components, and then performs a linear regression of the residuals on **X** is one of the many solutions existing for $\lambda = 0$. This algorithm is called principal component analysis (PCA) in the sequel, and it is similar to the correction method of Price et al. (2006) used in association studies.

Ridge Estimates

For $\lambda > 0$, the solution of the regularized least-squares problem is unique, and the corresponding matrices are called the "ridge estimates." The minimization algorithm starts with an SVD of the explanatory matrix, $\mathbf{X} = Q\Sigma\mathbf{R}^T$, where **Q** is an $n \times n$ unitary matrix, **R** is a $d \times d$ unitary matrix and $\boldsymbol{\Sigma}$ is an $n \times d$ matrix containing the singular values of **X**, denoted by $(\sigma_j)_{j=1..d}$. The ridge estimates are computed as follows

$$\hat{\mathbf{W}} = \mathbf{Q}\mathbf{D}_\lambda^{-1}\text{svd}_K(\mathbf{D}_\lambda\mathbf{Q}^T\mathbf{Y}) \quad (3)$$

$$\hat{\mathbf{B}}^T = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id}_d)^{-1}\mathbf{X}^T(\mathbf{Y} - \hat{\mathbf{W}}), \quad (4)$$

where $\text{svd}_K(\mathbf{A})$ is the rank K approximation of the matrix **A**, $\mathbf{Id}_d$ is the $d \times d$ identity matrix, and $\mathbf{D}_\lambda$ is the $n \times n$ diagonal matrix with coefficients defined as

$$\mathbf{d}_\lambda = \left(\sqrt{\frac{\lambda}{\lambda+\sigma_1^2}}, \ldots, \sqrt{\frac{\lambda}{\lambda+\sigma_d^2}}, 1, \ldots, 1\right).$$

A mathematical proof of this result is provided in a supplementary text, Supplementary Material online. The above equations describe a fast algorithm for computing the estimates of the parameter matrices **U**, **V**, **B**. The computational cost of this algorithm is mainly determined by the algorithmic complexity of the SVD. According to (Halko et al. 2011), computing the estimates requires $O(npK)$ operations. This complexity reduces to $O(np \log K)$ operations when random projections are used (our implementation). Accounting for the computational cost of $\mathbf{Q}^T\mathbf{Y}$, the complexity of the estimation algorithm is of order $O(n^2p + np \log K)$. For studies in which the number of samples, n, is much smaller than the number of response variables, p, the computing time of the ridge estimates is approximately the same as running the SVD algorithm on the response matrix twice.

Statistical Tests

Our new version of LFMM dissociates the estimation of latent factors from the tests of association with the primary (environmental) variables. To test association between the primary variables and each response variable, Y_j, we use the latent

score estimates obtained from the LFMM model as covariates in multivariate regression models. Those regression models evaluate the effects of the variables ($\mathbf{X}$) on the molecular markers and test the nullity of effect sizes. Suppose that a single primary variable is tested ($d = 1$, the extension to $d > 1$ variables is straightforward). We fit a multivariate linear regression model for each locus (ℓ)

$$\mathbf{y}_\ell = \mathbf{x}\beta_\ell + \hat{\mathbf{U}}\mathbf{v}_\ell^T + \mathbf{e}_\ell, \quad \ell = 1, \ldots, p, \tag{5}$$

where the K factors in $\hat{\mathbf{U}}$ are considered fixed and their corresponding effect sizes, $\mathbf{v}_\ell$, are then (re-)estimated. To test the null hypothesis $H_0 : \beta_\ell = 0$, we use a Student distribution with $n - K - 1$ degrees of freedom. To improve test calibration and false discovery rate (FDR) estimation, we eventually apply an empirical-null testing approach to the test statistics (Efron 2004).

Separating the estimation of latent factors from the testing phase has the advantage of allowing some flexibility when performing the tests. For example, including the latent factor estimates in tests based on generalized linear models, mixed linear models or robust linear models can be easily implemented in the LFMM framework. In the case of linear mixed models (LMM), the covariance matrix for random effects could be computed from the estimated factors as $\mathbf{C} = \hat{\mathbf{U}}\hat{\mathbf{U}}^T/n$ (Note that the mixed model terminology may sometimes be misleading. LMMs incorporate "fixed and random effects" whereas LFMMs incorporate "fixed and latent effects." Thus an LMM can use estimates computed by an LFMM.). In practice, we used the simple linear models which revealed themselves computationally efficient and performed well in simulations. Our two-step approach is similar to other methods for confounder adjustment in association studies (Price et al. 2006). It differs from the other approaches through the latent scores estimates, $\hat{\mathbf{U}}$, that, in our case, capture the part of response variation not explained by the primary variable. The methods presented in this study and their extensions are implemented in the R package lfmm.

Results

Simulation Study

In a first series of computer experiments, we compared the runtimes of the new version of LFMM to the former version used with its default parameter settings (LFMM 1.5, Frichot and François 2015). Several values of the number of individuals (n), markers (p), and number of factors (K) were simulated. The user runtimes for the Markov chain Monte Carlo algorithm implemented in LFMM 1.5 ranged between 8 min ($n = 100$, $p = 1{,}000$, $K = 2$) and 32.5 h ($n = 1{,}000$, $p = 20{,}000$, $K = 15$). Note that the results for LFMM 1.5 were obtained for a single CPU, and that the multi-threated version of the program runs significantly faster. With the same data sets and a single CPU, the user runtimes for LFMM 2.0 ranged between 0.5 s ($n = 100$, $p = 1{,}000$, $K = 2$) and 12.5 s ($n = 1{,}000$, $p = 20{,}000$, $K = 15$). The results represent an improvement of several orders compared with the previous version (fig. 1), meaning that much larger data sets could be analyzed with the new version within much shorter time lags (supplementary fig. S1, Supplementary Material online). For larger value of p, the relative difference between the two versions stabilized, and LFMM 2.0 ran 10,000–100,000 times faster than LFMM 1.5. Because strong effect sizes were simulated at causal markers, both versions had high power to detect those target markers (Supplementary fig. S2, Supplementary Material online). With these simulation parameter settings, the LFMM 2.0 tests had higher power and precision than those of LFMM 1.5.

In a second series of computer experiments, three "fast" association methods were applied to the simulated data: PCA (Price et al. 2006), Confounder Adjusted Testing and Estimation (CATE) (Wang et al. 2017) and our new version of LFMM (fig. 2, $n = 200$, $p = 10{,}000$). We compared the relative performances of the methods over 50 replicates by considering low to high intensities of confounding. The intensity of confounding corresponded to the percentage of variance of the variable of interest explained by the confounding factors in the simulated data. The runtimes of CATE and PCA were of the same order as LFMM 2.0. CATE was slower than LFMM 2.0 (for large K), and with our implementation of PCA using an improved SVD algorithm, PCA was faster than LFMM 2.0.

For lower values of confounding intensity, the three methods had small rates of false discoveries and high power to discover the target markers. For higher values of confounding intensity the performances of LFMM 2.0 were superior to the other methods and showed the best combination of power and FDR as measured by the F-score (fig. 2). Note that the lower performances of PCA were expected because this method does not use the variable of interest when estimating the hidden variables. Thus PCA does not exactly address the problem of estimating confounders, and has lower power than the other methods.

Humans and Climate

To detect genomic signatures of adaptation to climate in humans, we performed a GEA study using 5,397,214 SNPs for 1,409 individuals from the 1000 Genomes Project (2015), and bioclimatic data from the WorldClim database (Fick and Hijmans 2017). The size of the data sets represents one of the largest GEA study conducted so far. Nine confounders were estimated by a cross-validation approach. This estimated number was confirmed by the visual inspection of factor 9 showing more noise than information (fig. 3A, Supplementary figs. S3 and S4, Supplementary Material online). The factors mainly described correlation between population structure and climate in the sample, and differed from principal components of the genomic data. PCA, CATE, and two variants of LFMM 2.0 led to a list of 1,335 SNPs after pooling the list of candidates from the four methods (expected FDR = 5%, fig. 3B). A variant prediction analysis reported an over-representation of genic regions (665/1,335) with a large number of SNPs in intronic regions (fig. 3C). Top hits represented genomic regions important for adaptation of humans to bioclimatic conditions. The hits included functional variants in the *LCT* gene, and SNPs in the *EPAS1* and *OCA2* genes previously reported for their role in adaptation to

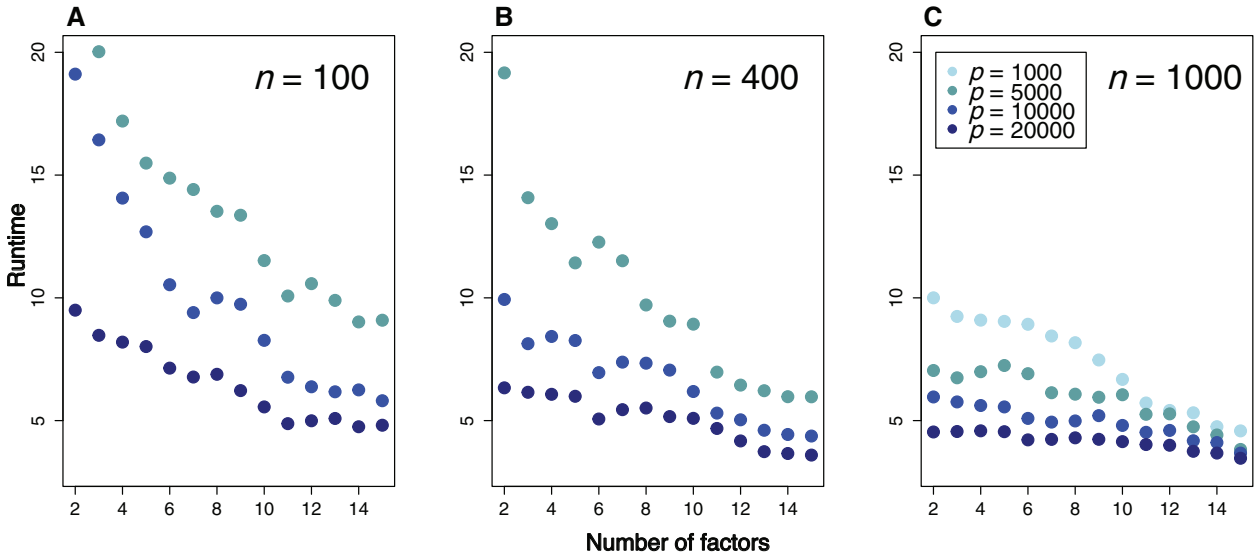

FIG. 1. Base 10 logarithm of the ratio of runtimes for LFMM 1.5 and LFMM 2.0. A value of 5 means that LFMM 2.0 runs 10^5 times faster than LFMM 1.5. (*A*) $n = 100$ individuals, (*B*) $n = 400$, (*C*) $n = 1{,}000$, p is the total number of markers in the simulation.

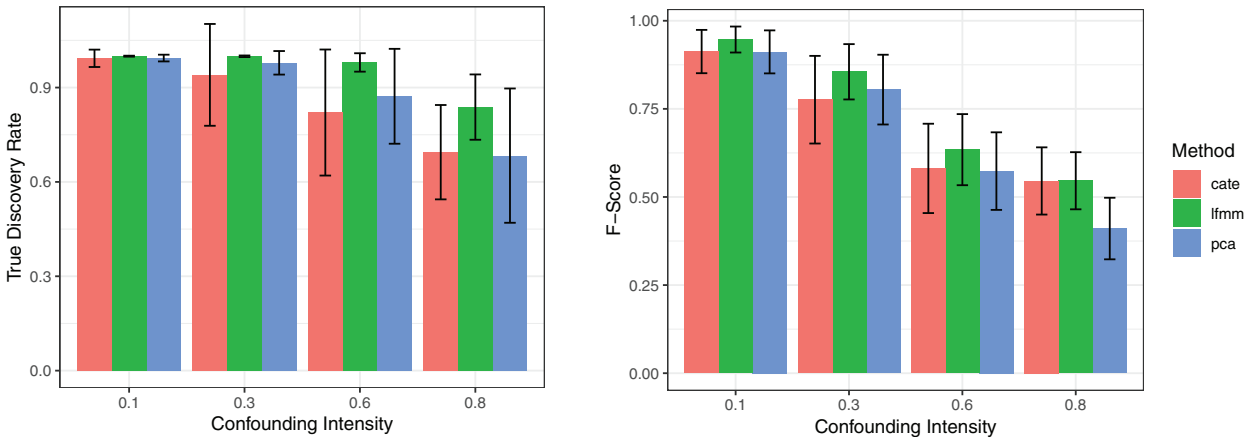

FIG. 2. True discovery rate and *F*-score as a function of confounding intensity. Three fast methods are considered: CATE, LFMM 2.0 and PCA. All methods were applied with $K = 8$ factors as determined by a PCA screeplot. The *F*-score is the harmonic mean of the true discovery rate (precision) and power.

diet, altitude or in eye color (Fan et al. 2016) (fig. 3*B*, Supplementary fig. S5, Supplementary Material online, and Supplementary table S1, Supplementary Material online).

RA and Smoking

Tobacco smoking is considered an established risk factor for the development of RA, an autoimmune inflammatory disease (Di Giuseppe et al. 2014). We performed an association study using whole blood methylation data from a study of patients with RA considering tobacco consumption as an environmental exposure variable (Liu et al. 2013). The goal of this study was to identify CpG sites exhibiting joint association with smoking and RA. The cell composition of blood in RA patients is a known source of confounding (Jaffe and Irizarry 2014), and we accounted for cell-type heterogeneity by using $K = 5$ factors in PCA, CATE, and LFMM 2.0. Like in a previous analysis, we combined the significance values of the three methods to increase power. The list of CpG sites showing significant joint association with RA and smoking in at least two approaches was short (nine CpG sites). The top-list included the genes *NMUR1* and *LYN* that play an important role in the regulation of innate and adaptive immune responses (fig. 4). *NMUR1* was identified as a hub gene in a protein–protein interaction network of differentially methylated genes in osteosarcoma, and its abnormal DNA methylation may contribute to the progression of the disease (Chen et al. 2018). The gene *LYN* acts downstream two genes related to RA and synovial sarcoma, *EPOR* and *KIT* (Tamborini et al. 2004; Huber et al. 2008; Kosmider et al. 2009), and it mediates the phosphorylation of *CBL* in relation to RA (Xu et al. 2018). This gene is also linked to *IL3* receptors associated to RA and smoking (Takano et al. 2004; Miyake et al. 2014). Regarding the next hits, *MPRIP* was found to be hypermethylated for patients with RA (Lin and Luo 2017), and association between

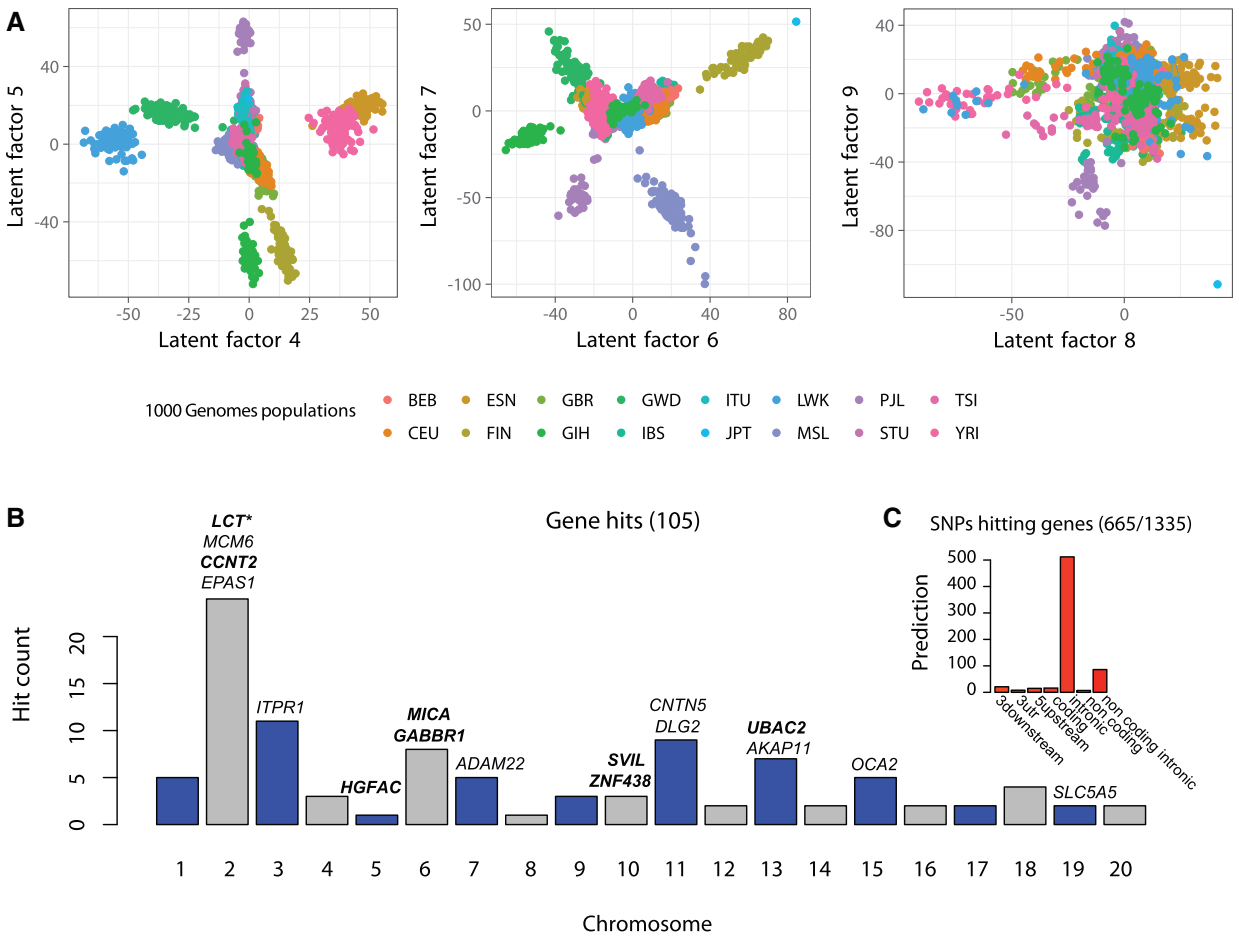

FIG. 3. Human GEA study. Association study based on genomic data from the 1000 Genomes Project database and climatic data from the Worldclim database. (*A*) Latent factors estimated by LFMM 2.0. (*B*) Target genes corresponding to top hits of the GEA analysis (expected FDR level of 5%). The highlighted genes correspond to functional variants. (*C*) Predictions obtained from the VEP program.

CXCR5 and RA or upregulation of this gene in the rheumatoid synovium has been reported in the literature (Schmutz et al. 2005; Wengner et al. 2007).

Discussion

In this study, we introduced LFMM 2.0, a fast and accurate algorithm for estimating confounding factors and for testing GEAs. The new algorithm is based on the exact solution of a regularized least-squares problem for latent factor regression models. We used LFMM 2.0 for testing associations between a response matrix and a primary variable matrix in a study of natural selection in humans. In addition, we used it to evaluate the importance of DNA methylation in modulating the effect of smoking in patients with an inflammatory disease. Previous inference methods for latent factor regression models were based on slower algorithms or on heuristic approaches, lacking theoretical guarantees for identifiability, numerical convergence, or statistical efficiency (Leek and Storey 2007; Wang et al. 2017). In addition, existing methods do not always address the confounding problem correctly, building their estimates on genetic markers only while ignoring the primary variables. For example, genome-wide association studies adjust for confounding by using the largest PCs of the genotypic data (Price et al. 2006). A drawback of the approach is that the largest PCs may also correlate with the primary variables, and removing their effects results in loss of statistical power. When compared with PCA approaches, LFMMs gained power by removing the part of genetic variation that could not be explained by the primary variables (Frichot et al. 2013). Thus LFMM extends the tests performed by the PCA approaches by improving principal components with factor estimates depending on the primary and response variables simultaneously.

In gene expression and DNA methylation studies where batch effects are source of unwanted variation, alternative approaches to the confounder problem have also been proposed. These methods are based on latent factor regression models called surrogate variable analysis (SVA) (Leek and Storey 2007; Wang et al. 2017). For epigenomic or gene-expression studies, LFMMs extend SVA and their recent developments implemented in CATE. Latent factor regression models employ deconvolution methods in which unobserved batch effects, ancestry or cell-type composition are integrated in the regression model using hidden factors. Those models have been additionally applied to transcriptome analysis (van Iterson et al. 2017). As they do not make specific hypotheses regarding the nature of the data, LFMMs and other latent

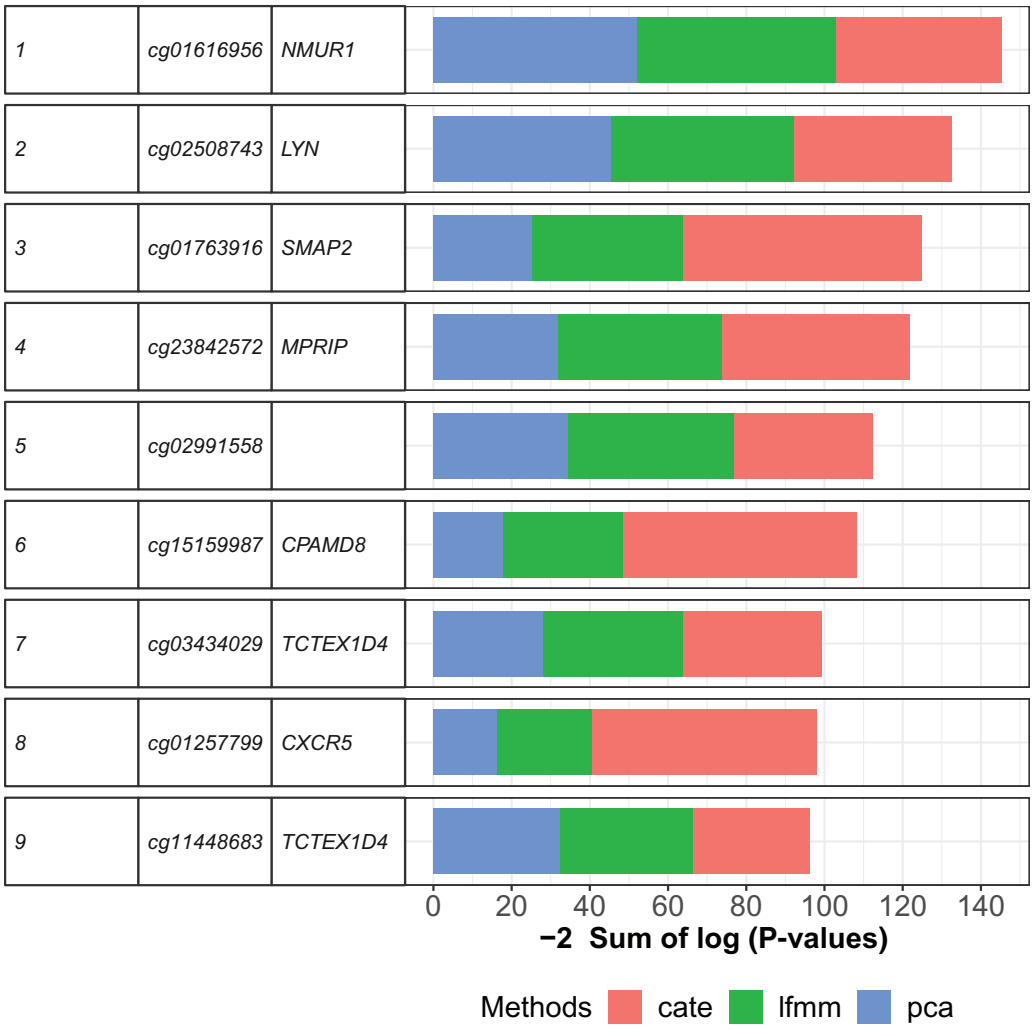

Fig. 4. EWAS of RA and smoking. Fisher's scores for CpG sites showing significant association with RA and smoking in at least two of three approaches (PCA, CATE, LFMM 2.0).

factor regression models could be applied to any category of association studies regardless of their application field.

Like several factor methods, the computational speed of LFMM methods is mainly influenced by the algorithmic complexity of low rank approximation of large matrices. The algorithmic complexity of the LFMM method was similar to PCA and CATE, of order $O(n^2p + np \log K)$ for LFMM 2.0. These approaches are much faster than the previous version of LFMM or than Bayesian methods currently used in GEAs.

Since the models underlying versions 1.5 and 2.0 of LFMM are alike, their statistical limitations are also similar. More specifically, estimation of latent factors might be complicated by physical linkage, unbalanced study designs or a strong correlation between axes of genetic variation and environmental gradients (McVean 2009; Frichot et al. 2015). Our new implementation of LFMM disconnects the testing steps from the latent factor estimation steps. This disconnection facilitates the implementation of approaches that alleviate the above issues. For example, the human SNP data were pruned for LD by taking the most informative SNPs in genomic windows before estimating latent factors. Although potential improvements such as sparse modeling, random effects, logistic or robust regressions and stepwise conditional tests were not included in our results, those options are available with the lfmm program, and they may provide additional power to detect true associations in GEA studies.

Materials and Methods

Simulation Data

We simulated primary variables, **X**, latent factors, **U**, and a response matrix $\mathbf{Y}_0$ according to a multivariate Gaussian model. In those simulations, we controlled the correlation between the primary variables and the confounders. More specifically, a primary variable, **X**, and three latent variables, **U**, were simulated by using a multivariate Gaussian distribution to represent individual data

$$(\mathbf{U}, \mathbf{X})_i \sim \mathrm{N}(0, \mathbf{S}),$$

where **S** was a covariance matrix with diagonal terms $(s_1^2, \ldots, s_K^2, 1)$, and nondiagonal terms set to zero, except for the covariance between $\mathbf{u}_k$ and **X**, which was set to the

value ρc_k. We created $K = 3$ confounders, assuming that their variances, (s_k^2), were equal to the values 15, 3, 0.1, respectively. The c_k coefficients were sampled from a uniform distribution taking values in the range $(-1, 1)$, and ρ was inversely proportional to the square root of $\sum_k c_k^2/s_k^2$ (which was < 1). The coefficient of proportionality was chosen so that the percentage of variance of **X** explained by the latent factors ranged between 0.1 and 1. The effect size matrix, **B**, was generated by setting a proportion of effect sizes to zero. Nonzero effect sizes were sampled according to a standard Gaussian distribution, $N(b, \sigma_b^2)$. The proportion of null effect sizes was set to 99%. We eventually created a response matrix, **Y**, by simulating $\mathbf{Y}_0$ from the generative model of the latent factor model, transforming the values through a probit transform, and generating genotypes according to a binomial distribution bin(2, π), where π resulted from the probit transform. Runtimes of LFMM 2.0 and LFMM 1.5 were measured on a Xeon W-2145 CPU (3.70 GHz). Both programs were used with their default parameters settings including 5 runs of 10,000 cycles for the Markov chain Monte Carlo algorithm in LFMM 1.5. In the programs K was varied in the range 2–15. To evaluate the capabilities of methods to identify true positives, we used the true discovery rate, power and the *F*-score. The true discovery rate (or precision) is the proportion of true positives in a candidate list of positive tests. Power is the number of true positives divided by the number of true associations. The *F*-score is the harmonic mean of the true discovery rate and power.

Other Algorithms

In the R programing environment, we considered the following methods and software: 1) We implemented a standard approach that estimates confounders from a PCA of the response matrix **Y**, and uses linear regression to perform the tests. Principal components were estimated with the svd function of the RSpectra package. Distinct scalings of the response matrix were used in simulations or in EWAS analysis and in the GEA analysis. For the GEA analysis, we used the scaling procedure implemented in the EIGENSTRAT method (Price et al. 2006), whereas in EWAS, we scaled with division by standard deviations. 2) We implemented the CATE method (Wang et al. 2017), which uses a linear transformation of the response matrix such that the first axis of this transformation is colinear to **X** and the other axes are orthogonal to **X**. One property of CATE is to use robust regression models to perform statistical testing. CATE was used without negative controls and with a test recalibration option similar to genomic control (Devlin and Roeder 1999). The CATE method was implemented in the R package cate. 3) We implemented a variant of LFMM 2.0 using an L^1 norm regularizer instead of the L^2 norm (Lasso estimates). All programs contained several options and many algorithmic variants. Unless specified, we used the default options of programs.

GEA Study

We performed a GEA study using whole genome sequencing data and bioclimatic variables to detect genomic signatures of adaptation to climate in humans. The data are publicly available, and they were downloaded from the 1000 Genomes Project (2015) and from the WorldClim database (Fick and Hijmans 2017). The genomic data included 84.4 millions of genetic variants genotyped for 2,506 individuals from 26 world-wide human populations. Nineteen bioclimatic data were downloaded for each individual geographic location, considering capital cities of their country of origin. The bioclimatic data were summarized by projection on their first principal component axis. The genotype matrix was preprocessed so that SNPs with minor allele frequency <5% and individuals with relatedness >8% were removed from the matrix. Admixed individuals from Afro-American and Afro-Caribbean populations were also removed from the data set. After those filtering steps, the response matrix contained 1,409 individuals and 5,397,214 SNPs. We performed LD pruning to retain SNPs with the highest frequency in windows of one hundred SNPs, and identified a subset of 296,948 informative SNPs. Four GEA methods were applied to the 1000 Genomes Project data set: PCA, CATE, and two LFMM estimation algorithms. For all methods the latent factors were estimated from the pruned genotypes, and association tests were performed for all 5,397,214 loci. Because the results were highly concordant, the significance values were combined by using the Fisher method. The results obtained from clumps with an expected FDR level of 1% were analyzed using the variant effect predictor (VEP) program (McLaren et al. 2016).

RA Data Set

The RA data are publicly available and were downloaded from the GEO database under the accession number GSE42861 (Liu et al. 2013). For this study, beta-normalized methylation levels at 485,577 probed CpG sites were measured for 354 cases and 335 controls (Liu et al. 2013). Following (Zou et al. 2014), probed CpG sites having a methylation level < 0.1 or >0.9 were filtered out. Two primary variables were included in the model, tobacco consumption and the health outcome. Ex-smokers were removed from the analysis, and all filtering steps resulted in 345,067 CpGs and 234 cases and 225 controls. Tobacco consumption was encoded as an ordinal variable with three levels (nonsmokers, occasional smokers, and regular smokers), and the health outcome was encoded as a dichotomous variable. Age and gender were included as nuisance variables. The goal of the study was to identify CpG sites with joint association with tobacco smoking and RA. The data were centered and scaled for a standard deviation of one. Since the cell composition of blood in RA patients typically differs from that in the general population, there is a risk for false discoveries that stem from unaccounted-for cell-type heterogeneity (Jaffe and Irizarry 2014). Since cell-type heterogeneity was not measured, we used latent factors to model it (Zou et al. 2014). Three methods were applied to the RA data set: PCA, CATE, and our new version of LFMM. The number of factors was set to $K = 5$ according to the screeplot of the PCA eigenvalues. For each method, significance values for the association with smoking and RA were combined using a squared-max transform that guaranteed that the resulting *P*-values follow a uniform distribution under the null hypothesis. Candidate lists of CpG

sites were obtained by using the Fisher method after correction for multiple testing with a 5% type I error.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors are grateful to two anonymous reviewers for their constructive comments. They thank M.G.B Blum, all organizers and participants of the SSMPG 2017 workshop held in Aussois for their feedback on the program. They also thank Katie Lotterhos and Matthieu Gautier for fruitful discussions during this workshop. This work was supported by a grant from LabEx PERSYVAL Lab, ANR-11-LABX-0025-01, and by a grant from French National Research Agency (Agence Nationale pour la Recherche) ETAPE, ANR-18-CE36-0005. This article was developed in the framework of the Grenoble Alpes Data Institute, supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

References

Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. 2008. High-dimensional sparse factor modeling: applications in gene expression genomics. *J Am Stat Assoc*. 103(484):1438–1456.

Chen XG, Ma L, Xu JX. 2018. Abnormal DNA methylation may contribute to the progression of osteosarcoma. *Mol Med Rep*. 17(1):193–199.

De Mita S, Thuillet AC, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y. 2013. Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Mol Ecol*. 22(5):1383–1399.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55(4):997–1004.

Di Giuseppe D, Discacciati A, Orsini N, Wolk A. 2014. Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. *Arthritis Res Ther*. 16(2):R61.

Eckart C, Young G. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* 1(3):211–218.

Efron B. 2004. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J Am Stat Assoc*. 99(465):96–104.

Fan S, Hansen ME, Lo Y, Tishkoff SA. 2016. Going global by adapting local: a review of recent human adaptation. *Science* 354(6308):54–59.

Fick SE, Hijmans RJ. 2017. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *Int J Climatol*. 37(12):4302–4315.

Frichot E, Schoville SD, Bouchard G, François O. 2013. Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol Biol Evol*. 30(7):1687–1699.

Frichot E, François O. 2015. LEA: an R package for landscape and ecological association studies. *Methods Ecol Evol*. 6(8):925–929.

Frichot E, Schoville SD, de Villemereuil P, Gaggiotti OE, François O. 2015. Detecting adaptive evolution based on association with ecological gradients: orientation matters! *Heredity* 115(1):22–28.

Gautier M. 2015. Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics* 201(4):1555–1579.

Günther T, Coop G. 2013. Robust identification of local adaptation from allele frequencies. *Genetics* 195(1):205–220.

Halko N, Martinsson PG, Tropp JA. 2011. Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev*. 53(2): 217–288.

Huber R, Hummert C, Gausmann U, Pohlers D, Koczan D, Guthke R, Kinne RW. 2008. Identification of intra-group, inter-individual, and gene-specific variances in mRNA expression profiles in the rheumatoid arthritis synovial membrane. *Arthritis Res Ther*. 10(4):R98.

Jaffe AE, Irizarry RA. 2014. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 15:R3.

Kosmider O, Buet D, Gallais I, Denis N, Moreau-Gachelin F. 2009. Erythropoietin down-regulates stem cell factor receptor (Kit) expression in the leukemic proerythroblast: role of Lyn kinase. *PLoS One* 4(5):e5721.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*. 3(9):e161.

Lin Y, Luo Z. 2017. Aberrant methylation patterns affect the molecular pathogenesis of rheumatoid arthritis. *Int Immunopharmacol*. 46:141–145.

Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. 2013. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*. 31(2):142–147.

Lotterhos KE, Whitlock MC. 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Mol Ecol*. 24(5):1031–1046.

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl variant effect predictor. *Genome Biol*. 17(1):122.

McVean G. 2009. A genealogical interpretation of principal components analysis. *PLoS Genet*. 5(10):e1000686.

Miyake Y, Tanaka K, Arakawa M. 2014. IL3 rs40401 polymorphism and interaction with smoking in risk of asthma in Japanese women: the Kyushu Okinawa maternal and child health study. *Scand J Immunol*. 79(6):410–414.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal component analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 38(8):904–909.

Rakyan VK, Down TA, Balding DJ, Beck S. 2011. Epigenome-wide association studies for common human diseases. *Nat Rev Genet*. 12(8):529–541.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R. 2015. A practical guide to environmental association analysis in landscape genomics. *Mol Ecol*. 24(17):4348–4370.

Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. *Nat Rev Genet*. 14(11):807–820.

Schmutz C, Hulme A, Burman A, Salmon M, Ashton B, Buckley C, Middleton J. 2005. Chemokine receptors in the rheumatoid synovium: upregulation of CXCR5. *Arthritis Res Ther*. 7(2):R217.

Takano H, Tomita T, Toyosaki-Maeda T, Maeda-Tanimura M, Tsuboi H, Takeuchi E, Kaneko M, Shi K, Takahi K, Myoui A, et al. 2004. Comparison of the activities of multinucleated bone-resorbing giant cells derived from CD14-positive cells in the synovial fluids of rheumatoid arthritis and osteoarthritis patients. *Rheumatology* 43(4):435–441.

Tamborini E, Bonadiman L, Greco A, Gronchi A, Riva C, Bertulli R, Casali PG, Pierotti MA, Pilotti S. 2004. Expression of ligand-activated KIT and platelet-derived growth factor receptor β tyrosine kinase receptors in synovial sarcoma. *Clin Cancer Res*. 10(3):938–943.

Teschendorff AE, Relton CL. 2018. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet*. 19(3):129.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

van Iterson M, van Zwet EW, Heijmans BT. 2017. Controlling bias and inflation in epigenome-and transcriptome-wide association studies using the empirical null distribution. *Genome Biol*. 18(1):19.

Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE. 2014. Genome scan methods against more complex models: when and how much should we trust them? *Mol Ecol*. 23(8):2006–2019.

Villemereuil P, Gaggiotti OE. 2015. A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods Ecol Evol*. 6:1248–1258.

Wang J, Zhao Q, Hastie T, Owen AB. 2017. Confounder adjustment in multiple hypothesis testing. *Ann Statist*. 45(5):1863–1894.

Wengner AM, Höpken UE, Petrow PK, Hartmann S, Schurigt U, Bräuer R, Lipp M. 2007. CXCR5—and CCR7—dependent lymphoid neogenesis in a murine model of chronic antigen-induced arthritis. *Arthritis Rheum*. 56(10): 3271–3283.

Xu XX, Bi JP, Ping L, Li P, Li F. 2018. A network pharmacology approach to determine the synergetic mechanisms of herb couple for treating rheumatic arthritis. *Drug Des Devel Ther*. 12:967.

Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J. 2014. Epigenome-wide association studies without the need for cell-type composition. *Nat Methods* 11(3): 309–311.