ANALYSE COMPARATIVE ET CARACTÉRISATION DES PATRONS PANGÉNOMIQUES DE MÉTHYLATION DE L'ADN PLACENTAIRE ET DE SANG DE CORDON OMBILICAL

par

Marika Groleau

Mémoire présenté au Département de biologie en vue
de l'obtention du grade de maître ès sciences (M. Sc.)

FACULTÉ DES SCIENCES
UNIVERSITÉ DE SHERBROOKE

Sherbrooke, Québec, Canada, septembre 2023

Le 19 septembre 2023

Le jury a accepté le mémoire de Marika Groleau dans sa version finale.

Membres du jury

Professeur Pierre-Étienne Jacques
Directeur de recherche
Département de biologie

Professeur Luigi Bouchard
Codirecteur de recherche
Département de biochimie et de génomique fonctionnelle

Nicolas Gévry
Évaluateur interne
Département de biologie

Professeur Luc Gaudreau
Président-rapporteur
Département de biologie

REMERCIEMENTS

Le paragraphe des remerciements en est un bien ingrat à écrire. En si peu de mots, on ne peut rendre justice (encore moins originalement), mais j'ose espérer que les personnes concernées ont ressenti toute la reconnaissance que j'ai pu avoir pour elles ces dernières années; admettons qu'à ce point-ci, ce serait quand même un peu tard pour réellement remercier celles et ceux qui doivent l'être. C'est toutefois très sincèrement que j'écris les remerciements qui suivent.

J'ai eu la chance incommensurable d'être accompagnée durant mes études supérieures par un comité exceptionnel de trois chercheurs aux expertises, approches et personnalités complémentaires. Pierre-Étienne, tu mènes ton laboratoire et tes mille et un projets de recherche (et autre) avec tout le dynamisme, la curiosité et la rigueur qu'on puisse espérer. Tu nous montres à naviguer l'univers parfois tortueux de la recherche avec stratégie et confiance, mais aussi franc-jeu, et face à nos défis, tu as l'oreille empathique et le conseil juste. À mon humble avis, tu as su unifier tes qualités de chercheur et d'humain, faisant de toi un directeur de recherche inspirant. Merci pour ton accompagnement et tes encouragements tout le long de mon parcours. Luigi, si les circonstances ont fait que ça aura pris plus de cinq ans avant que l'on se voit finalement en personne, tu as pourtant toujours été un co-directeur très présent, engagé et à l'écoute. Je te suis reconnaissante pour les maints conseils scientifiques, critiques constructives et encouragements que tu m'as prodigués ces dernières années. Marie-France, tu as peut-être le titre de conseillère sur mon comité, mais c'est évident que ton implication a largement dépassé ce rôle. Merci pour toutes les opportunités que tu m'as offertes, ainsi que pour les nombreuses réunions si riches en discussions, challenges et conseils.

Aux membres du laboratoire de Pierre-Étienne, merci pour votre humour, votre support et toutes les discussions si pertinentes (ou non) qui ont mis du soleil sur chaque journée.

À mon amour, Samuel, ton engagement et ton soutien au cours de ces dernières années hautes en émotions sont bien au-delà des remerciements. Je suis extrêmement choyée de cheminer à tes côtés et que l'on puisse compter l'un sur l'autre malgré les épreuves.

SOMMAIRE

La méthylation de l'ADN est un mécanisme clé de régulation adaptative de l'expression des gènes, laquelle est particulièrement dynamique dans les tissus embryonnaires. À cause de cette adaptabilité et du lien entre la santé à long terme de l'enfant et la méthylation de l'ADN établie tôt dans la vie, l'étude de cette dernière à la naissance à partir de tissus fœtaux est de plus en plus importante. Le sang de cordon ombilical ou le placenta sont deux exemples de tissus pouvant être facilement échantillonnés et présentant des avantages et inconvénients distincts. C'est principalement sur le placenta que se sont penchés les travaux présentés dans ce mémoire. Cet organe fœtal est responsable des échanges de nutriments et de déchets entre la mère et l'enfant en plus de sécréter des hormones influençant la physiologie maternelle et fœtale. Sans surprise, les dysfonctionnements placentaires ont été associés à diverses complications de la grossesse, mais, malgré son importance, la réponse placentaire à l'environnement reste mal comprise. Influencées à la fois par la génétique et les expositions environnementales, les marques épigénétiques telles que la méthylation de l'ADN jouent un rôle central dans notre compréhension de la biologie placentaire et leur étude pourrait permettre une meilleure compréhension de ses adaptations et des complications de grossesse. Les deux projets de recherche qui seront présentés visaient donc à caractériser le méthylome placentaire en contraste avec celui d'un autre tissu fœtal, ainsi qu'à cartographier à travers tout le génome les variants associés à la variation des niveaux de méthylation de l'ADN.

Le premier projet avait comme objectif de mettre en lumière certains des rôles physiologiques du placenta en comparant son profil de méthylation de l'ADN à celui d'un autre tissu fœtal, le sang de cordon ombilical. Une analyse comparative de la méthylation de l'ADN a été effectuée à l'échelle du génome à partir d'échantillons de placenta et de sang de cordon appariés et a permis d'identifier 4 912 régions différemment méthylées, lesquelles avaient principalement des niveaux plus bas dans le placenta. Des analyses de termes ontologiques réalisées sur le sous-ensemble de ces régions moins méthylées dans le placenta et situées en amont des sites d'initiation de la transcription ont montré un enrichissement pour les termes liés aux fonctions

des micro-ARN ainsi qu'aux gènes codant les récepteurs couplés aux protéines G. Ces résultats mettent en évidences des régions génomiques qui soulignent les fonctions du placenta.

Le deuxième projet avait comme objectif d'identifier et de caractériser les variants génétiques liés à la variation de la méthylation de l'ADN dans le placenta. Cette analyse visait à mieux comprendre la régulation de la méthylation de l'ADN dans ce tissus et ces résultats pourraient être utilisés pour bonifier des analyses portant sur la santé de l'enfant en lien avec son développement *in utero* et en bas âge. Il a été possible d'identifier 188 529 loci quantitatifs de méthylation de l'ADN, lesquels sont déplétés en îlots CpG et présentent un enrichissement pour des régions qui ne sont pas associées à un gène. Parmi ceux qui l'étaient, les analyses de termes ontologiques ont fait ressortir des fonctions connues du placenta, suggérant une spécificité tissulaire. Le chevauchement entre les régions différemment méthylées du premier projet et les résultats de ce deuxième projet suggéraient que des loci génétiques pourraient différemment réguler la méthylation de régions génomiques entre deux tissus, mais cette piste reste à être explorée dans des travaux futurs.

Mots-clés : méthylation de l'ADN, mQTL, DMR, épigénomique, placenta, sang de cordon ombilical, DOHaD, bio-informatique.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS

5caC	5-6carboxylcytosine
5fC	5-formylcytosine
5hmC	5-hydroxyméthylcytosine
ADN	Acide désoxyribonucléique
AFR	Ascendance africaine
AMR	Ascendance américaine
ARN	Acide ribonucléique
ARNm	ARN messager
BMI	*Body mass inde*x, indice de masse corporelle
CEU	Résidents de l'Utah d'ascendance nord et est-européenne
CGI	Îlots CpG
CHUS	Centre hospitalier universitaire de Sherbrooke
CpA	Dinucléotide cytosine–phosphate–adénine
CpC	Dinucléotide cytosine–phosphate–cytosine
CpG	Dinucléotide cytosine–phosphate–guanine
CpH	Dinucléotide composé d'une cytosine et de tout autre nucléotide qu'une guanine
CpT	Dinucléotide cytosine–phosphate– thymine
DNAm	*DNA methylation*, méthylation de l'ADN
DNMT	ADN méthyltransférase (de l'anglais *DNA methyltransferase*)
DOHaD	Origine développementale des maladies (de l'anglais *developmental origin of health and disease*)
EAS	Ascendance est-asiatique
eQTL	Locus de caractères quantitatifs d'expression génique
EUR	Ascendance européenne
EWAS	Analyse d'association panépigénomique (de l'anglais *epigenome-wide association study*)
FDR	*False discovery rate*, taux de faux positifs
Gen3G	*Genetics of Glucose regulation and Growth*
GO	*Gene ontology*

GRC	Genome Reference Consortium
GWAS	Analyse d'association pangénomique (de l'anglais *genome-wide association study*)
HGPO	Hyperglycémie provoquée par voie orale
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
mADN	Méthylation de l'ADN
MBD	Domaine de liaison des CpG méthylés (de l'anglais *methyl-CpG-binding domain*)
miRNA	*Micro RNA*, micro-ARN
mQTL	Locus de caractères quantitatifs de méthylation de l'ADN
OGTT	Test d'hyperglycémie provoquée (de l'anglais *oral glucose tolerance test*)
OR	*Olfactory receptor*, récepteur olfactif
pb	Paire de bases
PCR	Réaction en chaîne de la polymérase (de l'anglais *polymerase chain reaction*)
PolI, PolII et PolIII	ARN polymérase I, II et III
QTL	Locus de caractères quantitatifs (de l'anglais *quantitative trait loci*)
RNA-seq	Séquençage d'ARN
SNP	Polymorphisme nucléotidique (de l'anglais *single nucleotide polymorphism*)
tDMR	Région différemment méthylée entre deux tissus (de l'anglais *tissue-specific differentially methylated region*)
TET	Dioxygénase de methylcytosine à translocation dix-onze (de l'anglais *ten-eleven translocation methylcytosine dioxygenase)*
TPM	Transcrit par million
TSI	Toscans d'Italie
TSS200, TSS1500	Région génomique à moins de 200 ou 1500 nucléotides en aval du TSS
UTR	*Untranslated regions*, régions non traduites

CHAPITRE 1
INTRODUCTION

Le projet de recherche qui sera présenté dans ce mémoire porte sur l'analyse bio-informatique de l'interaction entre le génome et l'épigénome dans deux tissus fœtaux, le placenta et le sang de cordon ombilical. Avant d'en arriver à élaborer mes objectifs de recherche, il faut donc passer en revue les grandes lignes des deux domaines qui se rencontrent dans mon projet, la génomique et l'épigénomique, ainsi que l'incidence de l'un sur l'autre. Seront ensuite brièvement couverts les deux tissus fœtaux à l'étude ainsi que la cohorte humaine de laquelle proviennent les données ayant été utilisées. Finalement, les deux objectifs de recherche couverts par mes travaux de maîtrise seront énoncés.

1.1 La génomique

La génomique est le domaine de la biologie s'intéressant au génome entier d'un organisme, c'est-à-dire l'ensemble de ses gènes et régions intergéniques. Ce champ d'études est issu de l'élargissement du domaine de la génétique, qui elle s'intéresse seulement à des gènes spécifiques, ou un seul gène d'intérêt. La naissance de la génomique a probablement été propulsée par l'arrivée des technologies de séquençage à haut débit et par l'explosion de la capacité de calcul des ordinateurs modernes. De même, bien qu'il ne semble pas y avoir de nouveau terme en usage, la génomique a aussi évoluée en même temps que notre compréhension des génomes l'a fait. Ainsi, l'appellation génomique fait désormais référence à bien plus que l'étude de l'ensemble des gènes, mais aussi à celle des éléments régulateurs et structuraux de l'ADN.

En 1990, un grand projet ayant comme visée de séquencer tout le génome euchromatique humain a été lancé : *The Human Genome Project*, complété en 2004 (Abdellah *et al*., 2004). Outre la création de la première ébauche du génome humain, cet effort international a aussi mis en évidence une conception erronée du génome humain : alors que la communauté scientifique

pensait qu'il contenait entre 60K et 100K gènes codants pour des protéines (Fields *et al*., 1994), cette première construction semblait indiquer qu'il en contenait en fait ~20-25K, une faible surestimation par rapport à nos connaissances actuelles, soit ~19K (Piovesan *et al*., 2019). Cette première version présentait cependant de nombreuses lacunes : plus de 300 trouées dans les régions euchromatiques, des incertitudes quant à la taille de régions contenant des éléments répétés en plus de n'offrir aucune représentativité de la diversité génétique entre les différentes populations dans le monde (Church *et al*., 2011). Afin de s'attaquer à ces problèmes, mais aussi pour répondre au besoin d'uniformité et de stabilité dans la création de génomes de références, un nouveau consortium a été créé en 2007 : *The Genome Reference Consortium* (GRC). Depuis, c'est le GRC qui gouverne les versions et les mises à jour du génome humain, ainsi que ceux de la souris (*Mus musculus*), du poisson-zèbre (*Danio rerio*) et du coq (*Gallus gallus*).

Le plus récent assemblage du génome humain publié par le GRC, le GRCh38, se distingue des assemblages précédents entre autres par un plus grand nombre de loci alternatifs, soit des régions génomiques présentant une grande diversité génétique entre les populations humaines (Schneider *et al*., 2017). Par diversité génétique, on désigne toute variation entre des individus dans la séquence d'ADN. Les polymorphismes nucléotidiques (SNP), soit une substitution d'un seul nucléotide, sont les plus communément discutés, mais ne sont qu'un type parmi d'autres (par exemple, les insertions et délétions, les variants de nombre de copies, les inversions et les translocations). Le manque de représentation des diverses populations humaines dans les banques de données est actuellement une des plus grandes lacunes en génétique (Popejoy & Fullerton, 2016), soulignant l'importance d'efforts comme *The 1000 Genomes Project* (Auton *et al*., 2015) et *gnomAD* (Karczewski *et al*., 2020) pour offrir une meilleure connaissance de la diversité génétique humaine. D'ailleurs, une nouvelle version d'un génome de référence humain plus représentatif de la diversité génétique humaine a récemment été publié par le *Human Pangenome Reference Consortium* (Liao *et al*., 2023). Appelé le « pangénome », cette référence conçue sous forme de graphe permet la représentation simultanée de plusieurs variations du génome humain et inclura 350 individus d'ascendances distinctes en 2024.

1.1.1 Méthodes analytiques

Deux grandes familles de méthodes sont utilisées pour déterminer le génotype des individus : les méthodes basées sur les puces de génotypage, et celles basées sur le séquençage. Les puces de génotypage contiennent un certain nombre de sondes prédéfinies, qui correspondent aux positions que l'on souhaite génotyper. Parmi les puces les plus populaires, nommons la *Multi-Ethnic Genotyping Array* de Illumina, construite pour capturer la diversité génétique dans des cohortes aux ascendances multiples. Ces méthodes ont l'avantage d'être plus faciles et moins coûteuses, mais les positions qui ne sont pas mesurées doivent être imputées si on veut arriver à une aussi bonne couverture que le séquençage, ce qui ajoute une source d'erreur (Das *et al*., 2016). En ce qui concerne le séquençage, la technologie qui est encore la plus utilisée est celle d'Illumina (Heather & Chain, 2016). Si on souhaite pouvoir phaser les génotypes (en quelque sorte, reconstruire les chromosomes en assignant chacun des allèles à une des deux copies des chromosomes), le séquençage Illumina doit cependant être remplacé ou complété par des technologies capables de produire de longues séquences, comme celle de *PacBio*, *Oxford Nanopore* ou 10X (Pollard *et al*., 2018). Après le séquençage, les lectures doivent être alignées sur un génome de référence et des algorithmes doivent être utilisés afin d'assigner les génotypes (*SNP calling*) à chacune des positions. Pour cette dernière étape, c'est la suite GATK du *Broad Institute* qui établit actuellement les meilleures pratiques (DePristo *et al*., 2011).

1.2 L'épigénomique

Alors que la génomique, tel qu'il vient d'en être discuté, s'intéresse à la séquence même de l'ADN et particulièrement aux séquences encodant des gènes, l'épigénomique s'intéresse aux différentes modifications chimiques ciblant l'ADN qui, sans en changer la séquence ou le contenu informationnel intrinsèque, affectent l'expression des gènes et la programmation cellulaire. Ces modifications sont appelées des modifications épigénétiques et, ensemble, elles forment l'épigénome. Les marques épigénétiques sont passées d'une génération cellulaire à l'autre, mais aussi parfois d'une génération humaine à l'autre, et évoluent au fil de la vie et des expositions environnementales. Les sites du *National Human Genome Research Institute* (NIH,

2020) et de *MedlinePlus – National Library of Medicine* (NIH, 2021), deux offrent des définitions simples de l'épigénomique et de l'épigénétique, dont la distinction ne semble pas toujours faire consensus.

Les modifications épigénétiques les plus connues sont la méthylation des cytosines des séquences cytosines-guanine (CpG) de l'ADN (par simplicité, ultérieurement référée comme la méthylation de l'ADN, mADN) et les différentes modifications post-traductionnelles de queues d'histones, principalement la (poly)méthylation et l'acétylation des lysines (Figure 1.1). La mADN ainsi que les modifications d'histones seront discutées plus en détail plus loin. D'autres éléments mesurables en lien avec la chromatine sont souvent intégrés aux analyses ciblant l'épigénome. Par exemple, nommons le niveau d'accessibilité de la chromatine, les patrons de positionnement des nucléosomes, la présence de variants d'histones ou encore les profils d'expression des gènes. Ces différentes mesures ont en commun avec les marques épigénétiques d'être informatives sur l'état de la chromatine et donc, ultimement, sur l'expression des gènes.

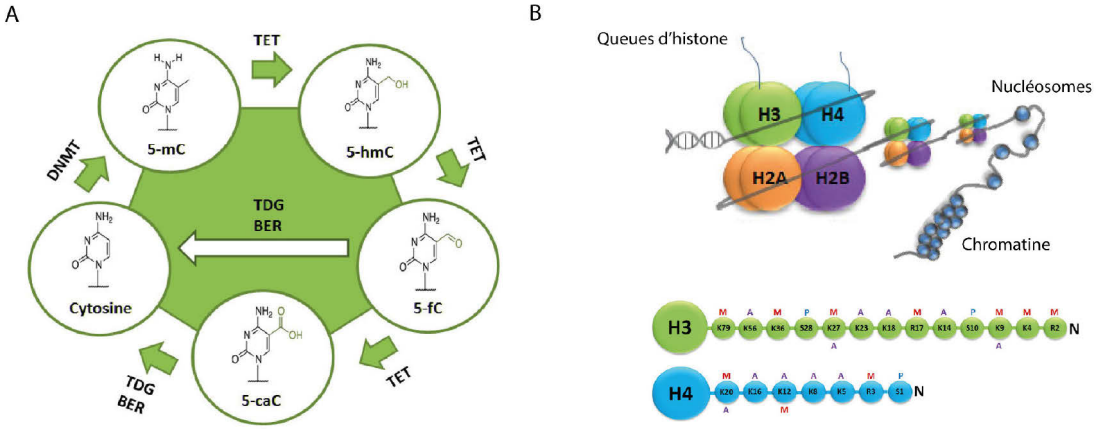

Figure 1.1 **Schématisation de la mADN et des modifications d'histones.**
A) Processus de méthylation et de déméthylation d'un CpG. Modifié de (WhatIsEpigenetics, 2013a). **B)** Les variants d'histones et leurs modifications les plus fréquentes. Modifié de (WhatIsEpigenetics, 2013b).

Par simplification et par manque d'outils et de connaissances, l'état de la chromatine a longtemps été catégorisé de façon binaire : l'hétérochromatine, sa forme condensée et inactive, et l'euchromatine, sa forme relaxée et disponible à la transcription. On parle encore en ces termes, mais la compréhension des différents états possibles de la chromatine, et de la nature des interactions intraépigénome sous-jacentes, a grandement évoluée. Des consortiums comme ENCODE (Moore *et al*., 2020), puis le IHEC (Bujold *et al*., 2016), ont vu le jour et permettent d'explorer les épigénomes de différents types cellulaires ainsi que les patrons des différentes marques épigénétiques qu'ils regroupent. Ces plateformes seront cruciales pour le futur de la biologie, car elles permettent de faire le pont entre les études plus ciblées (type cellulaire, marque épigénétique particulière, etc.) et les études intégratives. Les unes ne disposant pas de toute l'information permettant d'interpréter le mécanisme étudié et les autres manquant de support expérimental pour expliquer ces mécanismes, on comprend bien le rôle que de tels dépôts de données peuvent jouer.

1.2.1 La méthylation de l'ADN

Bien qu'à proprement dit la « méthylation de l'ADN » puisse référer à tout ajout de groupement méthyle sur une des bases constituant l'ADN, son usage en contexte eucaryote désigne presque assurément le 5e carbone des cytosines en contexte CpG. Il existe d'autres types de méthylation de l'ADN : les méthyladénines (fréquentes chez les procaryotes, mais également observées chez certains eucaryotes, dont l'humain (Xiao *et al*., 2018)) en plus des méthylcytosines dans les contextes CpA, CpT et CpC (ensemble nommées CpH) (Patil *et al*., 2014). Cependant, la méthylation des cytosines en contexte CpG est la modification la plus fréquente et la mieux caractérisée chez les eucaryotes, d'où l'abus de langage répandu dont ce document ne sera pas épargné.

1.2.1.1 Mise en place et maintien de la mADN

Les enzymes responsables de la méthylation de l'ADN sont nommées ADN méthyltransférases (DNMT, de leur nom anglais *DNA methyltransferase*). Il existe chez l'humain trois DNMT actives, DNMT1, DNMT3A et DNMT3B (Greenberg & Bourc'his, 2019). Les DNMT3A/B sont principalement responsables de la mise en place de la méthylation *de novo*, donc d'apposer la méthylation à des séquences qui en sont dépourvues sur les deux brins. Avec DNMT3L, elles sont cruciales à la mise en place de la mADN durant le développement embryonnaire et leur inhibition est létale chez la souris (Okano *et al*., 1999). Une fois la mADN mise en place, celle-ci doit généralement être maintenue au fil des divisions cellulaires. C'est ici qu'intervient une particularité de méthylation des CpG, qui la distingue de la méthylation des séquences CpH: par la nature symétrique de la séquence CpG (c'est-à-dire que sur le brin complémentaire on retrouve aussi la séquence CpG), l'ADN est presque toujours méthylé sur les deux brins. Quand ce n'est pas le cas, par exemple lors de la réplication de l'ADN, on parle alors d'hémiméthylation. C'est DNMT1 qui, grâce à son recrutement aux sites hémiméthylés aux fourches de réplication, s'occupe du maintien des patrons de mADN sur le brin nouvellement synthétisé (Sharif *et al*., 2007).

1.2.1.2 Modification de la mADN

La déméthylation de l'ADN peut survenir soit par un mécanisme passif, soit par un mécanisme actif. La voie passive consiste en un défaut de reméthylation suivant la réplication de l'ADN, phénomène par exemple observé chez l'embryon humain entre le stade deux cellules et le stade du blastocyste (Greenberg & Bourc'his, 2019). La voie active est quant à elle dépendante des enzymes TET (de l'anglais *ten-eleven translocation methylcytosine dioxygenase*). Les TET oxydent de façon itérative les méthylcytosines en 5-hydroxyméthylcytosine (5hmC), en 5-formylcytosine (5fC) puis en 5-6carboxylcytosine (5caC) (Figure 1.1). Ces formes de cytosines modifiées pourraient être reconnues par d'autres protéines possédant un domaine de liaison à l'ADN et avoir leurs propres fonctions épigénétiques, bien que ce ne soit pas encore clair (Kohli & Zhang, 2013) (Figure 1.2). Le processus de déméthylation actif peut ensuite lui-même

prendre deux voies : la déméthylation « passive » par la réplication de l'ADN, puisque les formes oxydées des méthylcytosines ne permettent pas le recrutement de DNMT1, ou le remplacement actif des 5fC et 5caC par des cytosines via les voies de réparation par excision de base (Kohli & Zhang, 2013). Il n'existe pas d'enzyme capable de directement retirer le groupement méthyle des méthylcytosines.

1.2.1.3 Interprètes de la mADN

Avant de passer au survol des rôles de la mADN, il est pertinent d'avoir une idée des protéines capables de reconnaître les (méthyl)cytosines. Deux domaines protéiques semblent avoir une importance particulière pour cette tâche : le domaine CXXC, qui reconnaît les dinucléotides CpG non méthylés (Xu *et al*., 2018) et le domaine MBD (methyl-CpG-binding domain), qui reconnaît les dinucléotides CpG méthylés (Q. Du *et al*., 2015). Dans les deux cas, les protéines connues pour posséder un de ces domaines (ou les deux) participent presque toutes à un processus de modification de la chromatine (par exemple, la méthylation / déméthylation de l'ADN ou la méthylation / acétylation d'histones). Outre les protéines possédant des domaines spécifiques de reconnaissance des (methyl)-CpG, d'autres protéines pouvant lier l'ADN, notamment des facteurs de transcription, peuvent présenter plus ou moins d'affinité de liaison à leur motif de reconnaissance si ce dernier comporte des méthylcytosines (Greenberg & Bourc'his, 2019).

1.2.1.4 Rôles de la mADN

Chez les mammifères, 70-80% des CpG sont méthylés (Li & Zhang, 2014), mais cette méthylation n'est pas équitablement répartie dans le génome, entre autres parce que, sans coïncidence, les CpG eux-mêmes ne sont pas distribués uniformément. Les méthylcytosines étant plus propices à la désamination spontanée, qui mène à une transition C en T, le génome des espèces où la mADN est fréquente est appauvri en CpG (Cooper & Krawczak, 1989). On observe cependant certaines régions qui, au contraire, sont denses en dinucléotides CpG (les « îlots CpG », ou CGI) et que ces îlots sont fréquemment déméthylés et associés à des promoteurs

de gènes, à proximité des sites d'initiation de la transcription (TSS, de l'anglais *transcription start site*) (Deaton & Bird, 2011) (Figure 1.2). Considérant cela, et que plus de 50% des méthylcytosines se situent dans des régions génomiques au rôle inconnu chez l'humain (Greenberg & Bourc'his, 2019), il semblerait qu'une grande partie de la mADN n'ait pas d'importance fonctionnelle (Edwards *et al*., 2017). Malgré cela, les rôles connus qui lui sont attribués sont critiques au bon fonctionnement cellulaire.

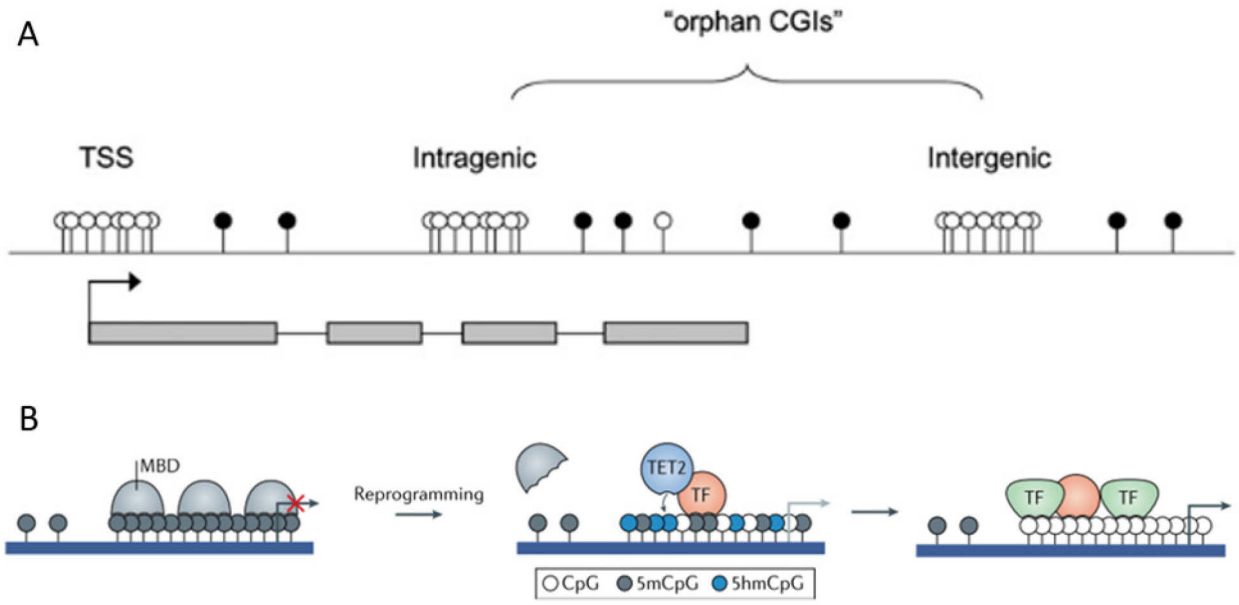

Figure 1.2 **Exemples de positionnement et de rôles de la mADN**

A) Exemple d'un promoteur associé à un CGI, de CGI non associés à un promoteur (« orphan CGI ») et de CpG non associés à des îlots. **B)** Exemple de reprogrammation épigénétique. Un promoteur est méthylé et occupé par des protéines liant les méthylcytosines, mais l'arrivée d'un facteur de transcription liant préférentiellement les méthylcytosines permet le recrutement de TET2 et de rendre la chromatine accessible pour d'autres facteurs de transcription. Modifié de Deaton *et al.* (2011) (A) et de Greenberg & Bourc'his (2019) (B).

De façon générale, on peut considérer la mADN comme une marque épigénétique répressive, c'est-à-dire qu'elle est associée au l'inhibition de l'expression des gènes et aux régions génomiques silencieuses. Une des fonctions probablement les plus anciennes de la mADN est de protéger le génome contre l'ADN exogène, ce qui se traduit chez l'humain par l'inactivation des rétrotransposons encore actifs (Bourc'his & Bestor, 2004). La mADN joue aussi un rôle dans le maintien de la répression à long terme des gènes, notamment ceux ciblés par l'inactivation aléatoire d'un des deux chromosomes X chez les femelles (Gendrel *et al*., 2012) ainsi que ceux soumis à l'empreinte parentale (Monk *et al*., 2019). En dehors de ces contextes génomiques spécifiques, la relation entre la mADN et l'expression des gènes n'est pas toujours aussi claire.

Sans entrer dans les détails mécanistiques, il semblerait que les promoteurs associés à des CGI déméthylés (~70% des promoteurs humains (Long *et al*., 2013)) présentent une conformation de la chromatine qui est plus accessible à la machinerie de transcription (Deaton et Bird, 2011). L'absence de méthylation aux promoteurs CGI n'en garantit pas l'expression, qui serait plutôt régulée par d'autres mécanismes (par exemple, par les complexes répressifs Polycomb (Marasca *et al*., 2018)), mais la présence de méthylation à ces promoteurs est quant à elle indicatrice de la répression de ces gènes. Pour les promoteurs pauvres en CpG, la méthylation serait peu susceptible d'être un mécanisme régulateur de l'expression et la corrélation inverse pouvant être observée entre la mADN à ces promoteurs et l'expression génique pourrait plutôt être expliquée par l'activité transcriptionnelle qui protégerait les CpG contre l'action des DNMT (Edwards *et al*., 2017). Pour le corps des gènes, c'est toutefois le contraire qui est observé : le corps des gènes activement transcrits est enrichi en méthylcytosines (Ball *et al*., 2009). Cet enrichissement pourrait avoir comme rôle de réduire l'activité de promoteurs intragéniques cryptiques, ou encore être impliqué dans la régulation de l'épissage co-transcriptionnel (Shayevitch *et al*., 2018).

1.2.1.5 Méthodes d'analyse de la mADN

Il existe deux approches largement utilisées pour mesurer la mADN lorsque l'on s'intéresse à son statut à travers tout le génome, soit l'utilisation de puces et le séquençage. Ces deux classes de technologies ont été dérivées de celles déjà existantes et bien rodées pour le génotypage, mais avec certaines adaptations leur permettant, bien entendu, de distinguer les méthylcytosines des cytosines. La première étape distinctive, sur laquelle repose tout le reste, consiste à traiter l'ADN au bisulfite de sodium. Ceci a pour effet de convertir les cytosines en uracile en catalysant leur désamination, une réaction dont sont protégées les méthylcytosines. L'ADN converti au bisulfite de sodium est ensuite amplifié par une réaction en chaîne de la polymérase (PCR, de l'anglais *polymerase chain reaction*), et les uraciles remplacés par des thymines. Les puces à ADN servant à mesurer la mADN contiennent des sondes ciblant un ensemble de CpG d'intérêt dont le niveau de méthylation est mesuré de manière analogue au génotypage par puce. Le « gold standard » des puces de mADN chez l'humain est présentement la *Infinium MethylationEPIC BeadChip* d'Illumina, qui mesure plus de 850k sites CpG à travers le génome. Si l'ADN est plutôt séquencé après sa conversion bisulfite de sodium, les méthodes de séquençage régulières discutées à la section 1.1.1 peuvent être utilisées. Cependant, comme la complexité du génome est grandement diminuée en raison de la conversion de la plupart des cytosines en thymines, des algorithmes d'alignement adaptés doivent être utilisés. L'approche du séquençage est plus coûteuse et probablement plus complexe que l'utilisation de puces, mais permet d'avoir une couverture presque entière du génome, sans biais introduit par le choix de conception de la puce.

1.2.1.6 Représentation numérique de la mADN

À l'échelle d'un site CpG, l'état de la mADN est dichotomique : méthylé ou non-méthylé. Comme mentionné précédemment, ce site CpG peut cependant ne pas être méthylé sur les deux brins d'un même chromosome; on parle dans ce cas d'hémiméthylation. Si on prend maintenant en considération les deux copies du chromosome, on obtient quatre versions d'une même position pouvant être ou non méthylée. Si la méthylation de ce site, dans cette cellule, est

mesurée par une puce (laquelle ne peut distinguer ni le brin ni la copie du chromosome), on obtiendra une mesure de la méthylation qui sera exprimée en un taux de mADN. Pour en ajouter à la complexité, ces analyses sont généralement effectuées à partir d'un mélange cellulaire hétérogène. Ainsi, bien que la mADN soit une information dichotomique, elle est régulièrement exprimée sous forme de valeur continue représentant une tendance dans l'échantillon utilisé pour l'analyse. La perte de l'information granulaire peut être un inconvénient dans certains contextes d'analyses qui ne seront pas discutés ici, mais elle permet en contrepartie l'utilisation des modèles statistiques pour variables continues.

Plus concrètement, la mADN est typiquement exprimée soit en valeur bêta, soit en valeur M (Pidsley *et al*., 2013). Les valeurs bêta sont bornées entre 0 et 1 et peuvent être interprétées comme un pourcentage de méthylation. Elles tirent leur nom de la distribution homonyme, laquelle est la plus adéquate pour représenter le comportement aléatoire de variables représentant un taux ou un pourcentage. Malgré cela, la plupart des méthodes statistiques utilisées pour analyser les données de mADN exprimées en valeurs bêta se basent sur la loi normale. Cette pratique cause parfois problème, car la distribution de mADN en un site répond rarement aux prémices de l'utilisation des méthodes basées sur la loi normale, entre autres en ce qui concerne l'homoscédasticité, c'est-à-dire la constance de la variance sur l'étendue des valeurs. Pour répondre à cette limitation, certains auteurs utilisent plutôt les valeurs M, qui correspondent à une transformation logit des valeurs bêta (Formule 1.1) (P. Du *et al*., 2010).

$$M = log_2(\frac{\beta}{1-\beta}) \qquad (1.1)$$

La nouvelle distribution n'est alors pas bornée, allant de l'infini négatif à l'infini positif en passant par zéro, lequel représente approximativement 50% de méthylation. Cette transformation rend généralement les données plus adaptées à l'utilisation de modèles basés sur la loi normale, mais ce gain se fait au détriment de la facilité d'interprétation de la valeur puisqu'elle est moins intuitive à comprendre qu'un pourcentage de méthylation.

1.2.2 Les modifications post-traductionnelles d'histones

L'ADN des eucaryotes subit plusieurs niveaux de compaction, dont le premier est son enroulement autour d'octamères d'histones (Felsenfeld et Groudine, 2003). Retrouvées tout le long des chromosomes, ces structures autour desquelles l'ADN est enroulé sur 146 paires de bases (pb) se nomment les nucléosomes. Quatre types d'histones sont assemblés pour former en former le cœur : H2A, H2B, H3 et H4, tous représentés deux fois. Ces assemblages constituent le fondement de la chromatine et stabilisent l'ADN en une forme plus compacte; ils sont cependant loin d'être statiques. Dépendamment du contexte et des besoins cellulaires, les histones peuvent subir différentes modifications post-traductionnelles ou encore être remplacées par des variants aux fonctions spécifiques (Zhou *et al*., 2011). Bien que cette composante épigénétique ne soit pas concernée par les travaux présentés dans ce mémoire, son importance biologique lui mérite cette courte section.

Les modifications d'histones les mieux étudiées et dont l'importance est la plus claire sont sans doute l'acétylation et la méthylation de lysines des queues N-terminales des histones H3 (Figure 1.1). Généralement, l'acétylation des lysines d'histones est associée à des formes de la chromatine plus accessibles, car l'acétylation des lysines en neutralise la charge positive, ce qui déstabilise l'interaction des histones avec l'ADN, lui chargé négativement. En ce qui concerne la méthylation des lysines d'histones, son mécanisme d'action serait plutôt de changer la capacité des interacteurs des histones à les reconnaître. Dépendamment de la lysine particulière qui est modifiée, du contexte génomique ainsi que des autres modifications en co-occurrence, la méthylation des histones H3 peut être associée à différents états de la chromatine. La Figure 1.3, tirée de la revue de Zhou *et al.* (2011), résume une vision simplifiée de quelques modifications d'histones et des états de la chromatine auxquels elles sont associées en fonction du contexte génomique. L'ajout et le retrait de ces deux groupements aux queues des histones sont des processus actifs et adaptatifs dont sont responsables des enzymes dédiées nommées histones acétyltransférases, déacétylases (Voss et Thomas, 2018), méthyltranférases et déméthylases (Hyun *et al*., 2017).

En plus d'être la cible de modifications post-traductionnelles, les histones peuvent aussi être remplacées aux nucléosomes par des variants exprimés à différents moments du cycle cellulaire (Henikoff et Smith, 2015) afin de répondre aux besoins particuliers de la chromatine à un état cellulaire donné. Par exemple, l'histone H3 est remplacée par CENP-A aux centromères, qui joue un rôle important dans la formation des kinétochores (Henikoff et Smith, 2015), et est remplacé par H3.3 aux régions génomiques où les nucléosomes sont particulièrement dynamiques, par exemple en étant transcriptionnellement actifs (Orsi *et al*., 2009). Au-delà des fonctions spécifiques respectives de chacun de ces variants d'histone, il est intéressant de noter que le remplacement des histones aux nucléosomes, que ce soit par un de leurs variants ou par une histone équivalente, a également l'effet non négligeable de « réinitialiser » les modifications post-traductionnelles.

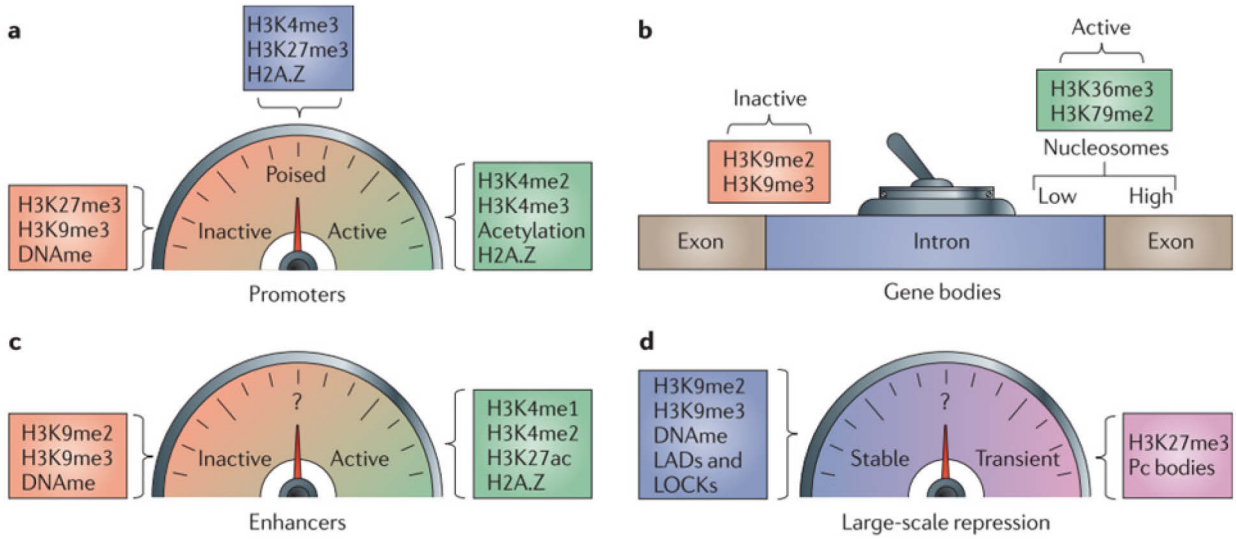

Figure 1.3 **Résumé des modifications d'histones les plus fréquentes et des états chromatiniens leur étant le plus associé en fonction du contexte génomique.**

Modifié de de Zhou *et al.* (2011).

1.2.3 La transcriptomique

La transcriptomique se définit comme l'étude de l'ensemble ou d'une partie des transcrits présents dans un contexte cellulaire donné. Comme abordé plus haut, la transcriptomique et l'épigénomique sont des compléments naturels l'un de l'autre. Si la transcriptomique permet de comprendre les impacts des modifications épigénétiques sur l'expression des gènes, l'épigénomique permet de remonter à la source des changements observés dans le transcriptome (par ex.: des changements dans l'abondance des transcrits ou la présence d'isoformes). En ce sens, quelques mots sur la transcriptomique sont aussi de mise.

D'abord, rappelons qu'il existe chez les animaux trois grandes familles d'enzymes responsables de la transcription de l'ADN, les ARN polymérases I, II et III (par abus de langage, PolI, PolII et PolIII pour la suite), identifiés depuis les années '70 (Roeder et Rutter, 1969). Ces complexes se distinguent par leurs gènes cibles ainsi que pas leurs mécanismes de régulation. La PolI est responsable de la transcription de l'ARN précurseur des grosses sous-unités ribosomales et la PolIII, de la petite sous-unité ribosomale ainsi que des ARN de transfert. C'est la PolII qui est alors responsable de la transcription des ARN messagers (ARNm), des micro ARN, ainsi que d'autres types d'ARN non codants (Dergai et Hernandez, 2019).

À toutes les étapes du processus de transcription (reconnaissance des cibles, initiation, élongation et terminaison), les polymérases sont dépendantes d'un ensemble de facteurs. En plus d'être requis pour la transcription, ces derniers sont nécessaires à la réalisation des événements co-transcriptionnels. Un modèle en « foyers » (*hub*) d'initiation ou d'élongation, par exemple, permet d'expliquer comment les polymérases changent de partenaires d'interactions et transitent d'une étape à l'autre (Figure 1.4). Quoiqu'ils ne puissent pas être discutés ici, notons que les polymérases sont entre autres associées à des complexes de remodelage de la chromatine (dont des modificateurs d'histones), ainsi qu'à des facteurs d'usinage (*processing*) des ARN (Cramer, 2019).

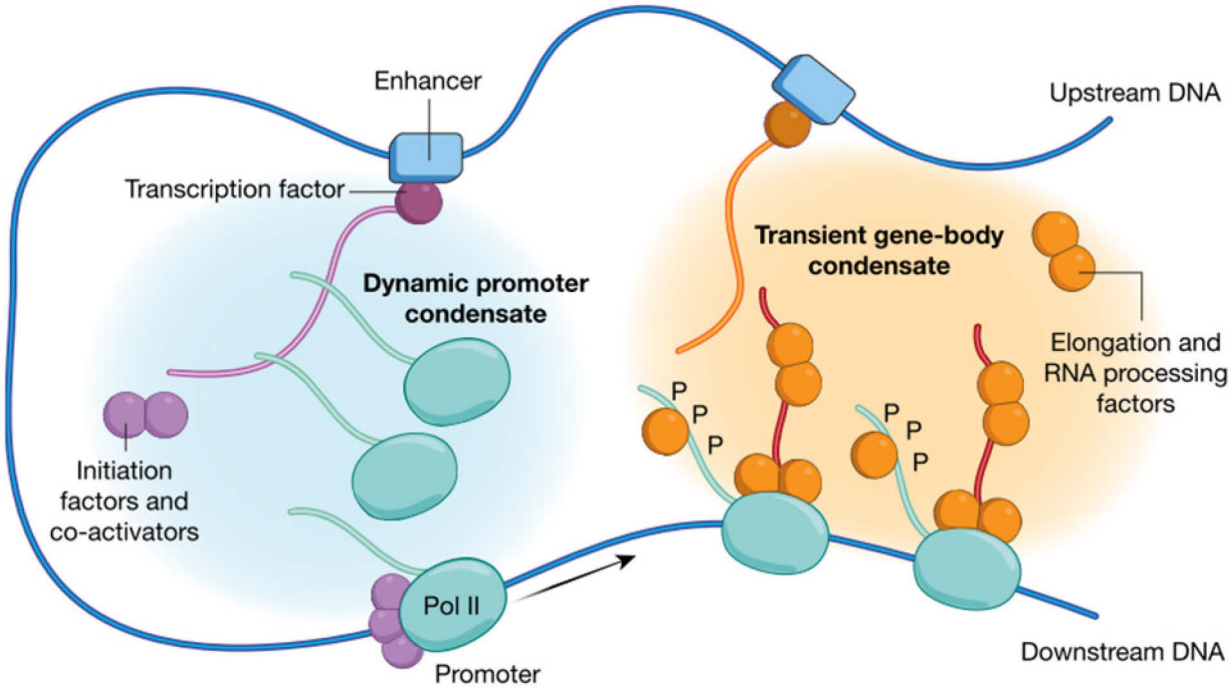

Figure 1.4 **Schématisation de foyers d'initiation et d'élongation pour la PolII.** Modifié de Cramer (2019).

1.2.3.1 Méthodes analytiques

D'un point de vue méthodologique, la transcriptomique pose quelques défis de plus que d'autres « omics », en partie à cause de l'instabilité des ARN et des biais qui peuvent être introduits par la façon dont les ARN sont sélectionnés. Pendant longtemps, les puces de quantification d'ARN étaient la méthode par défaut pour les études de transcriptomique, mais la famille de méthodes la plus utilisée de nos jours est le séquençage d'ARN (*RNA-seq*), les gains informationnels de cette méthode supplantant désormais son coût (Hrdlickova *et al*., 2017). Brièvement, l'ARN est extrait, les espèces d'ARN d'intérêt sont isolées puis rétrotranscrites et les fragments d'ADN ainsi produits sont envoyés pour séquençage. Chacune des étapes comporte plusieurs variantes, mais il faut apprécier l'importance du choix de la méthode de sélection des espèces d'ARN

(souvent, on cherche entre autres à éliminer les ARN ribosomaux qui constituent 80-95% des transcrits) et de la possibilité ou non de connaître le brin dont provient chacun des transcrits (Deluca *et al*., 2012; Hrdlickova *et al*., 2017).

1.2.4 Les interactions intraépigénomiques

Jusqu'à présent, les modifications épigénétiques ont été présentées de façon relativement indépendante, mais la réalité est toute autre. Les différentes modifications d'histones et la mADN sont grandement interdépendantes et le portrait complet de l'état de la chromatine pour une région (c'est-à-dire, sa densité en nucléosomes, les variants et modifications d'histones, la mADN, la présence d'ARN polymérase, et ainsi de suite) est un prédicteur nettement plus juste qu'une seule de ces mesures individuelles. C'est parce que c'est la combinaison spécifique des différentes marques épigénétiques qui donne à la chromatine ses caractéristiques, pas seulement leur addition indépendante (Adsera *et al*., 2019). La Figure 1.5 illustre trois exemples simples d'interactions entre les modifications d'histones et les DNMT. Plusieurs mécanismes épigénétiques complexes dépendent de cascades d'interactions plus sophistiquées afin de maintenir un état répressif de la chromatine, notamment les gènes soumis à l'empreinte parentale, le chromosome X inactivé chez les femelles, les gènes de régulation du développement ou encore des répresseurs de tumeur dans certains cancers (Deaton et Bird, 2011).

L'analyse intégrative de nombreuses marques épigénétiques a permis la catégorisation par (Ernst et Kellis, 2010) de certaines combinaisons de ces marques en 51 états chromatiniens distincts. Cet effort intégratif soulève l'importance d'interpréter les marques épigénétiques dans leur contexte global lorsque possible, puisqu'une même marque peut être associée à différents états de la chromatine.

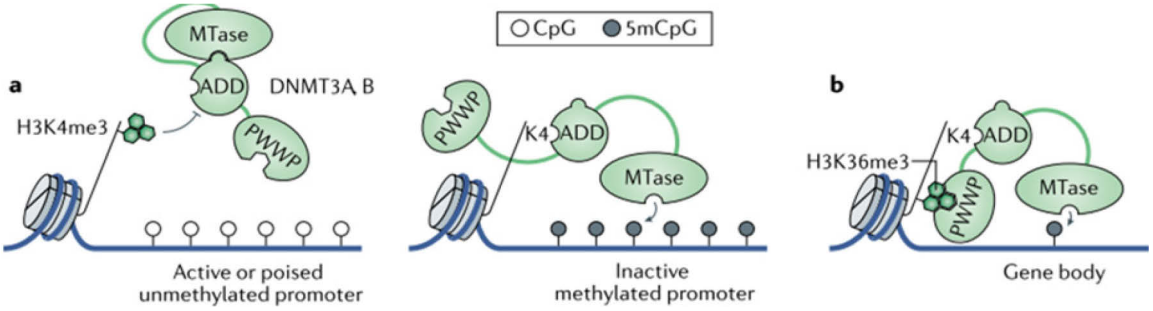

Figure 1.5 **Exemples d'interactions intraépigénome.**

A) Les promoteurs actifs sont enrichis en H4K4me3, ce qui empêche DNMT3A/B d'être recruté à l'ADN et de la méthyler (gauche). Même en cas de remplacement des histones H3 dans une telle région active, l'état d'activation est maintenu, car l'histone méthyltransférase de H3K4 possède un domaine CXXC qui reconnaît les cytosines non méthylées. Par contre, si la marque H3K4me3 s'avérait absente d'un tel promoteur, DNMT3A/B pourrait alors être recruté et méthyler l'ADN, ce qui empêcherait l'histone méthyltransférase de H3K4 d'être recruté par ce mécanisme (droite). **B)** Le corps des gènes activement transcrits est enrichi en H3K36me3 et déplété en H4K4me3, ce qui permet le recrutement de DNMT3A/B et la méthylation de l'ADN. Modifié de Greenberg et Bourc'his (2019).

1.3 Les interactions génome-épigénome-phénotype

La section précédente a couvert les interactions épigénomiques et son incidence sur l'expression des gènes. À la lumière de ces connaissances, il apparaît assez logique qu'il existe également d'importantes interactions entre les phénotypes et l'épigénome. De plus, tout comme il est assez bien connu que le génome ait une incidence sur les phénotypes, il en a aussi une sur les marques épigénomiques (Aguet *et al*., 2020; Cavalli et Heard, 2019). Il devient alors naturel de s'intéresser aux cas où le génotype est associé à la fois à une marque épigénétique et à un phénotype. C'est ce champ de recherche que la présente section survolera en abordant les notions de loci quantitatifs et d'inférence causale.

1.3.1 Les loci quantitatifs

Les loci de caractères quantitatifs (QTL, de l'anglais *quantitative trait loci*) sont des régions génomiques dans lesquelles les variants génétiques sont associées à la variation d'un trait mesurable. Les QTL sont typiquement détectés à l'aide de SNP qui, en étant corrélés avec la variation du trait d'intérêt, servent de représentant pour l'association causale qui, elle, demeure souvent cryptique. Il existe autant de types de QTL que de classes de traits mesurables, mais les plus connus sont les QTL associés à l'expression des gènes (eQTL), à la mADN (mQTL) ou à l'épissage des transcrits (sQTL, de l'anglais *splicing*) (Ye *et al*., 2020). L'identification de QTL sert principalement deux objectifs : un fondamental, la compréhension de la régulation du trait mesuré, et l'autre plus appliqué, la description des mécanismes derrière des traits complexes et les maladies (Figure 1.6). Des portails de données comme *eQTLdb* (Kerimov *et al*., 2020) et GTEx (Aguet *et al*., 2020) permettent de partager et d'accéder aux QTL identifiés dans différents tissus.

Les QTL trouvent une grande partie de leur utilité lors de l'interprétation de résultats d'études d'associations pangénomiques (GWAS, de l'anglais *genome-wide association study*) ou panépigénomiques (EWAS, de l'anglais *epigenome-wide association study*). En effet, quand des associations sont identifiées entre des SNP situés dans des régions intergéniques et un phénotype, l'explication biologique est rarement triviale. Grâce aux mQTL, il est parfois possible d'apporter une explication biologique à l'association trouvée dans un GWAS si le SNP en question est également impliqué dans un mQTL, et si le CpG du mQTL présente un intérêt dans le phénotype étudié. De façon complémentaire, les mQTL peuvent parfois être utilisés pour déterminer la direction d'une corrélation qui serait mesurée entre la mADN et un trait d'intérêt, type d'analyse appelée « inférence causale ».

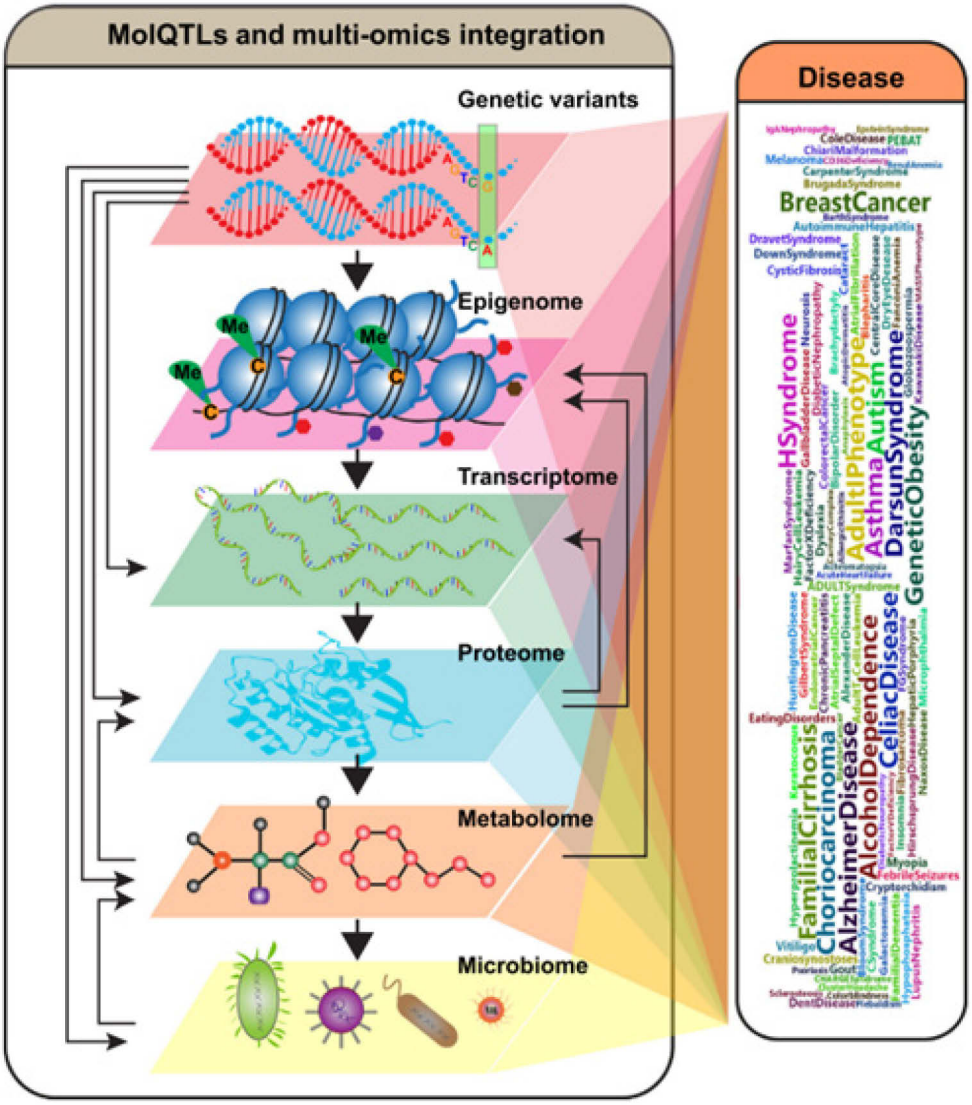

Figure 1.6 **Représentation des différentes couches « omics » pouvant être affectées par le génotype et de leurs relations entre elles, et avec des maladies potentielles**.

Modifié de Ye *et al.* (2020).

1.3.2 L'inférence causale

Lorsque l'on s'intéresse à un trait complexe, en plus de tenter d'en comprendre les processus physiopathologiques, on souhaite généralement identifier des facteurs protecteurs ou de risque pour ce trait afin d'en prédire le développement ou la présentation. Pour qu'un tel facteur soit fiable et utilisable, il ne faut pas simplement démontrer sa corrélation avec le phénotype, mais bien prouver qu'il existe une relation de causalité entre le facteur (cause) et le phénotype (effet). Cet exercice est plus épineux qu'il n'y paraît, comme en témoignent quelques exemples à la Figure 1.7. Notamment, il faut s'assurer du sens de la relation de causalité et que l'association soit libre de toute variable confondante, c'est-à-dire une variable qui affecterait à la fois le facteur et le phénotype (Pingault *et al*., 2018). La randomisation mendélienne est une famille de méthodes permettant d'inférer une telle causalité et d'en minimiser les pièges (G. D. Smith et Hemani, 2014). Brièvement, la randomisation mendélienne utilise les variants génétiques (typiquement des SNP) comme variable instrumentale. Ces dernières sont des variables mesurables, indépendantes de toute influence et corrélées avec le facteur d'intérêt (par ex.: l'expression d'un gène, la mADN) afin de servir de proxy pour estimer l'effet de ce facteur sur un phénotype. À titre d'exemple, supposons que l'on a trouvé une association entre le niveau de méthylation d'un CpG et une maladie et que l'on voudrait inférer la causalité. Il serait alors possible d'utiliser la randomisation mendélienne à condition de posséder un mQTL pour ce CpG, et que le SNP de ce mQTL soit aussi associé à la maladie. Cette application illustre bien la valeur de la découverte et du partage des mQTL.

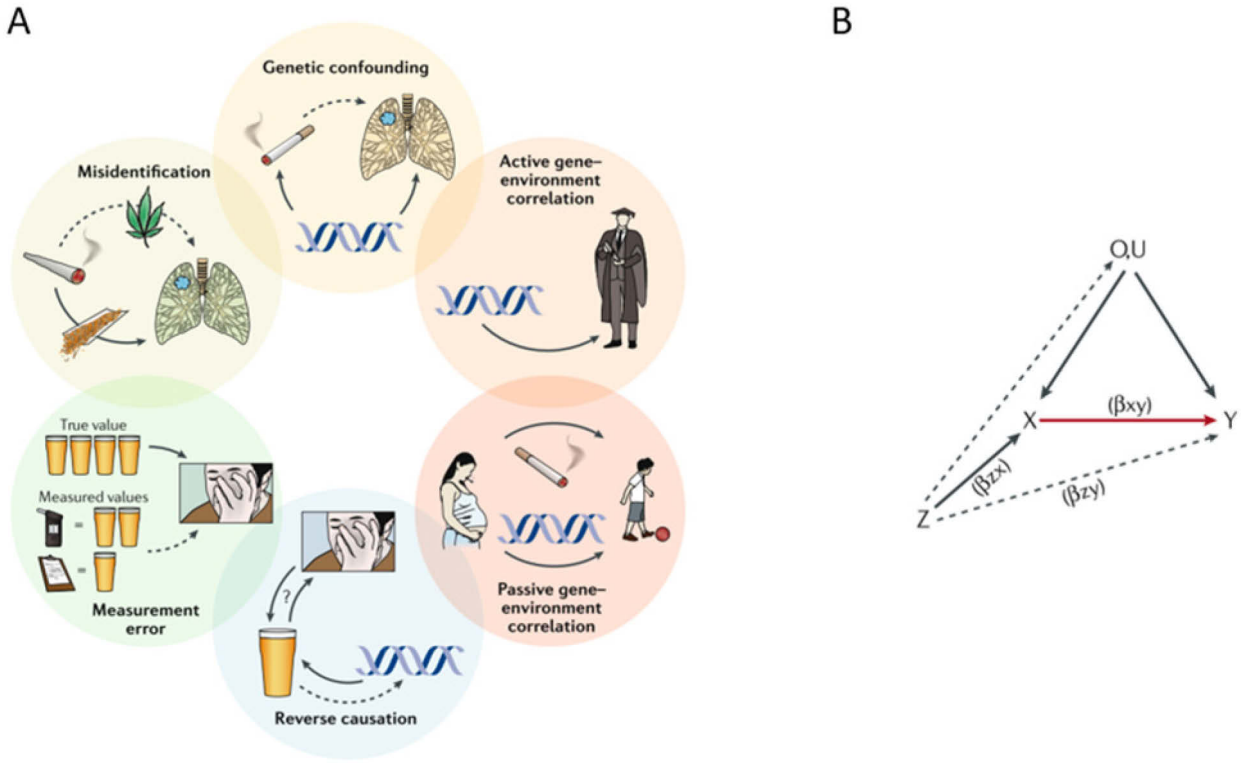

Figure 1.7 **Schématisation de concepts liés à l'inférence causale.**

A) Exemples de problèmes rencontrés lors de l'inférence de causalité. Par exemple, le génotype peut affecter l'environnement en prédisposant à l'atteinte d'un plus haut niveau de scolarité, ou encore être un effet confondant dans la relation entre la cigarette et le cancer du poumon (un SNP augmenterait le risque de cancer indépendamment du statut de fumeur, mais prédisposerait aussi à être un fumeur). **B)** Modèle de la randomisation mendélienne. Le génotype Z a un effet sur X, et un effet sur Y qui est entièrement médié par X. L'environnement (O,U) peut avoir des effets sur X et Y, mais Z doit en être complètement indépendant pour pouvoir inférer la causalité entre X et Y. Les lignes pointillées représentent des associations directes qui violeraient les hypothèses de la randomisation mendélienne. Modifié de Pingault *et al.* (2018).

1.3.3 Origine développementale des maladies

Développée dans les années '90, l'origine développementale des maladies (DOHaD, de l'anglais *developmental origin of health and disease*) est la théorie selon laquelle certaines maladies sont causées par des mésadaptations physiologiques en réponse à l'environnement fœtal ou néonatal (Barker, 1988). Ces adaptations à des facteurs environnementaux désuets (ou stress passagers) se feraient via des modifications épigénomiques, lesquelles peuvent persister dans le temps. L'exemple des enfants nés durant la famine néerlandaise de 1944 est fréquemment employé pour illustrer la DOHaD. En effet, il a été observé que les personnes dont les mères ont souffert de la famine durant la grossesse ont développé plus de problèmes cardiovasculaires et liés à l'obésité une fois adulte, ce qui s'explique par une adaptation à la carence en nourriture menant leurs organismes à accumuler davantage; une mésadaptation majeure dans des conditions d'abondance (Schulz, 2010). Ainsi, on a supposé que d'autres maladies ou traits complexes trouvent leurs origines (ou, minimalement, facteurs de risque) dans des expositions de la mère ou du nouveau-né, et cette hypothèse a été vérifiée maintes fois depuis (Ozanne et Constância, 2007). C'est pourquoi de nombreuses cohortes longitudinales récoltant des données sur les habitudes de vie de la mère, sur l'enfant à la naissance ainsi qu'au courant de sa vie ont été mises en place (Hoffman *et al*., 2017). Parmi les marqueurs d'intérêt mesurés chez l'enfant, notons la mADN de sang de cordon ombilical et de placenta (Breton *et al*., 2017).

1.4 Tissus biologiques d'intérêt pour ce mémoire

Comme énoncé plus haut, les données que j'utiliserai proviennent de deux tissus : le placenta et le sang de cordon ombilical. Puisque le placenta et le sang de cordon sont des tissus fréquemment utilisés en biologie du développement, mais plus rarement rencontrés hors de ces sphères, et qu'ils présentent des caractéristiques d'importance substantielle pour les analyses, quelques mots sur ceux-ci s'imposent.

1.4.1 Le placenta

Le placenta est l'organe responsable des échanges de nutriments entre la mère et l'enfant durant la grossesse, en plus d'exercer des rôles endocrines (Maltepe et Fisher, 2015). En tant que tel, le tissu placentaire est d'origine fœtale, donc génétiquement identique aux cellules somatiques de l'enfant; toutefois, il provient des cellules du trophectoderme, ce qui le distingue très tôt épigénétiquement de l'embryon lui-même (Koukoura *et al*., 2012). Afin d'accomplir ses fonctions d'échange, le placenta doit être en contact très étroit avec la circulation sanguine maternelle, ce qui est permis par l'invasion de cellules placentaire dans la paroi utérine. Le placenta humain, en forme de disque, crée ainsi une « poche » entre lui et la paroi utérine, l'espace intervilleux, qui est rempli de sang maternel et des villosités du placenta (Figure 1.8, (Maltepe et Fisher, 2015)). Les côtés fœtal et maternel du placenta diffèrent conséquemment en structure (donc également en composition cellulaire), ainsi qu'en fonction.

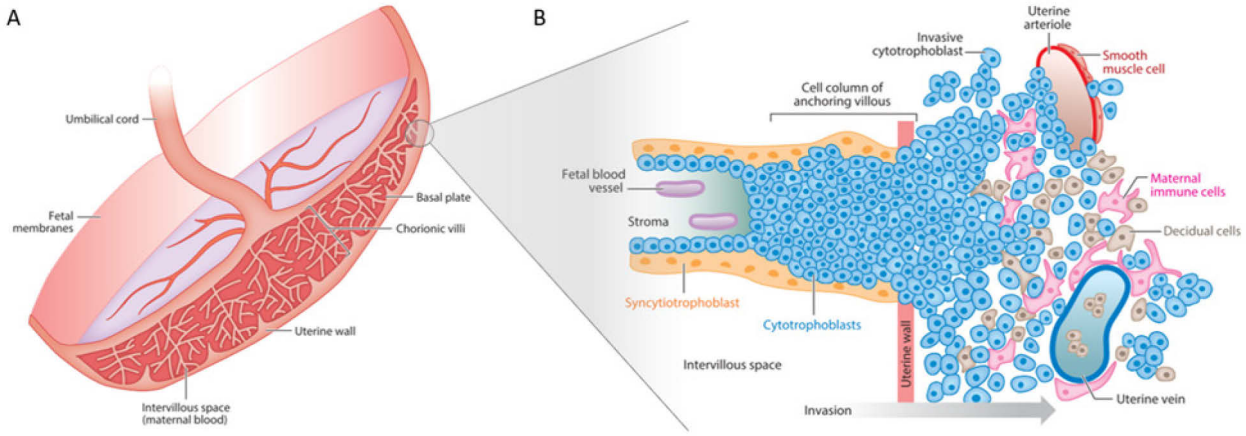

Figure 1.8 **Le placenta humain.**

A) Structure du placenta humain. **B)** Représentation de l'ancrage des vili placentaires dans la paroi utérine. Modifié de Maltepe et Fisher (2015).

1.4.1.1 La mADN dans le placenta

Le génome placentaire est connu pour présenter de plus faibles taux de mADN que les autres tissus ou types cellulaires sains (Bianco-Miotto *et al*., 2016). Cette hypométhylation n'est cependant pas observée de manière uniforme; ce sont plutôt certaines régions spécifiques du génome qui ont de faibles taux de méthylation, alors que le reste présente des niveaux de méthylation comparables à d'autres tissus. Ces régions de faible mADN sont nommées des domaines de méthylation partielle (Schroeder *et al*., 2013), et ceux spécifiques au placenta couvrent des gènes aux fonctions spécifiques à ce tissu et au processus développementaux.

1.4.2 Le sang de cordon ombilical

Le sang de cordon ombilical est quant à lui prélevé par ponction sanguine du cordon ombilical, et correspond au sang circulant chez le nouveau-né. Les défis liés à l'hétérogénéité cellulaire des échantillons de sang de cordon sont moins grands que ceux pour le placenta, entre autres grâce à de robustes méthodes d'estimation de sa composition cellulaire (Cardenas *et al*., 2016). Il présente en revanche un type cellulaire absent du sang adulte, ce qui complexifie les comparaisons : les érythroblastes, ou globules rouges nucléés. Cet aspect devra être pris en compte dans l'analyse des données de mADN provenant du sang de cordon, mais n'affecte pas les données de séquençage de génome entier produites à partir de ces prélèvements.

1.5 Cohorte prospective Gen3G

Mon projet de recherche s'inscrit dans le contexte de la cohorte prospective mère-enfant Gen3G (*Genetics of Glucose regulation and Growth*), qui a pour objectif d'approfondir la compréhension de la pathophysiologie du diabète gestationnel, ainsi que des mécanismes moléculaires liant le diabète gestationnel et l'obésité juvénile (Guillemette *et al*., 2016). Gen3G a initialement recruté plus de 1000 femmes qui ont été suivies tout au long de la grossesse, et amassé des données à la naissance de plus de 850 nouveau-nés au Centre Hospitalier Universitaire de Sherbrooke (CHUS). Des données physiologiques et concernant les habitudes

de vie ont été récoltées sur les participantes et des échantillons de placenta et de sang de cordon ombilical ont été prélevés au moment de l'accouchement. Lorsque possible, les mères et leur enfant ont aussi été rencontrées lors de suivis à 3 et 5 ans, à l'occasion desquels des échantillons sanguins ont entre autres été récoltés. Les nombreux projets découlant de Gen3G ont donné lieu à la mise en place d'une riche collection de données génétiques, épigénétiques, métabolomiques et de transcriptomique. L'intérêt d'étudier des tissus fœtaux, et particulièrement leur méthylome, réside dans le grand dynamisme de la mADN durant le développement, ce qui les rend plus sensibles aux influences de l'environnement (Bianco-Miotto *et al*., 2016). Bien que la mADN soit une modification réversible et changeant avec le vieillissement (Horvath, 2013), un patron épigénétique établi durant le développement peut aussi persister toute la vie d'un individu et affecter sa santé à long terme (Hoffman *et al*., 2017), comme le décrit la section 1.3.3 concernant l'origine développementale des maladies.

1.6 Questions de recherche

1.6.1 Contexte

Une des forces à exploiter dans les jeux de données de Gen3G est d'avoir plusieurs couches d'information pour un même individu. Même si toutes les expériences n'ont pas été menées sur le même sous-ensemble de participants (car, à l'origine, chacune répondait à des questions de recherche relativement indépendantes), le recoupement de ces jeux de données est généralement assez intéressant pour permettre des analyses de données intégratives. Ceci permet la formulation de nombreuses et riches questions de recherche. Dans le cadre de ma maîtrise, deux projets en continuité ont été proposés : comparer les patrons de mADN du placenta et du sang de cordon ombilical ainsi qu'identifier des mQTL pour la face fœtale du placenta. Bien que les objectifs spécifiques de ces deux projets soient distincts, ils servent tous deux des buts scientifiques communs. D'une part, ils permettent de contribuer à la caractérisation du méthylome placentaire en identifiant de nouvelles régions impliquées dans la biologie du placenta et de valider certains éléments de la littérature dans nos propres données. L'étude du placenta est un champ d'intérêt plutôt jeune qui s'avère complexe, particulièrement en ce qui a

trait à son épigénome. Ainsi, il est important que différents groupes se penchent encore sur la question en employant des méthodes complémentaires. D'autre part, en contribuant à la caractérisation du méthylome placentaire, ces projets faciliteront la recherche dont le but est d'étudier le rôle de cet organe durant la grossesse ainsi que son implication dans le développement de complications gestationnelles, comme le diabète gestationnel ou la prééclampsie.

1.6.2 Objectif 1 : Analyse comparative des patrons pangénomiques de méthylation de l'ADN du placenta et du sang de cordon ombilical

Dans le cadre de ce premier objectif, je me suis intéressée à la comparaison des profils de mADN dans le placenta et le sang de cordon. L'hypothèse de ce projet était qu'une analyse comparative entre le méthylome du placenta et du sang de cordon permettrait de mettre en évidence des régions abritant des gènes supportant de nouvelles fonctions du placenta. Ainsi, cette approche rendrait possible la validation des régions déjà décrites pour présenter un profil de mADN distinctif dans le placenta et pourrait suggérer de nouvelles régions intéressantes. De plus, cette analyse établirait une comparaison de la mADN à l'échelle du génome entre deux tissus fœtaux utilisés dans l'étude de la santé de la mère et de l'enfant. Les régions méthylées de façon similaire ou différente ainsi identifiées pourraient servir de référence pour faciliter la mise en relation des résultats d'études faites dans chacun des tissus. Les résultats de ce projet ont été publiés en 2021 et seront présentés au Chapitre 2 sous forme d'article scientifique.

1.6.3 Objectif 2 : Cartographie pangénomique des loci quantitatifs de méthylation de l'ADN dans le placenta

Pour mon deuxième objectif de maîtrise, je souhaitais adopter une approche davantage intégrative en ajoutant une dimension génétique à mes analyses. Ainsi, je me suis intéressée à la relation entre les variants génétiques et les mesures de mADN issues d'échantillons de la face fœtale du placenta. Plus précisément, cet aspect du projet consistait à cartographier les mQTL proximaux dans le placenta, puis à en faire une analyse descriptive (par exemple, en rapportant

leur relation aux CGI et leur position par rapport aux éléments régulateurs et gènes connus) ainsi qu'à les comparer avec ceux déjà présents dans la littérature. Ces résultats pourraient ensuite être ajoutés à des banques de données déjà existantes pour servir de référence, et être utilisés par d'autres groupes. D'ailleurs, le dernier volet de mon objectif deux était de tenter d'utiliser les mQTL pour bonifier des analyses déjà réalisées à l'intérieur ou à l'extérieur de la cohorte Gen3G. Ces travaux, non publiés, seront présentés sous forme classique au Chapitre 3.

CHAPITRE 2
ANALYSE COMPARATIVE DES PATRONS PANGÉNOMIQUES DE MÉTHYLATION DE L'ADN DU PLACENTA ET DU SANG DE CORDON OMBILICAL

2.1 Contexte de publication de l'article

L'article a été publié en ligne en accès libre le 4 mars 2021 dans le journal *Epigenomics* de l'éditeur *Future Medecine* (Groleau *et al*., 2021).

2.1.1 Contribution de l'article à la science

Dans ce projet, nous avons généré des données de mADN placentaires et de sang de cordon ombilical pour 444 échantillons appariés avec la puce EPIC d'Illumina, dans le but mettre en évidence les fonctions placentaires à partir des régions différemment méthylées entre les deux tissus (tDMR). Au meilleur de ma connaissance, de telles approches systématiques ont surtout été utilisées pour comparer les patrons de méthylation de tissus cancéreux et sains. Ainsi, nous montrons à partir de tissus sains qu'il est possible d'identifier des régions déjà connues pour être importantes dans le placenta en plus de mettre en lumière d'autres régions d'intérêt qui ont potentiellement un rôle intrinsèque dans la biologie du placenta. De plus, peu d'études disposent du même volume de données que Gen3G pour étudier le méthylome placentaire, ce qui ajoute à l'intérêt de cette contribution. En somme, ce projet se distingue par la méthode employée pour caractériser la mADN placentaire, autant par la quantité de données que par le modèle d'analyse différentielle avec un autre tissu fœtal sain apparié. De plus, nos résultats soulignent des régions où l'ADN est faiblement méthylé dans le placenta qui pourraient contribuer à comprendre la physiologie unique de cet organe, telles que des régions associées à des micro ARN, des récepteurs couplé aux protéines G ainsi que des récepteurs olfactifs.

2.1.2 Apport des auteurs

Marika Groleau (M.G.) a effectué les analyses, interprété les résultats et rédigé le manuscrit avec l'aide de Marie-France Hivert (M-F.H.), Luigi Bouchard (L.B.) et Pierre-Étienne Jacques (P-E.J.). Frédérique White et Andres Cardenas ont contribué aux analyses. Patrice Perron, M-F.H. et L.B. ont contribué à la mise sur pied et au financement de la cohorte, incluant la collecte des échantillons et des données. M.G., M-F.H., L.B. et P-E.J. ont contribué à l'élaboration du projet. Tous les auteurs ont révisé le manuscrit et approuvé sa version finale.

2.2 Comparative epigenome-wide analysis highlights placenta-specific differentially methylated regions

Marika Groleau[1], Frédérique White[1], Andres Cardenas[2], Patrice Perron[3,4], Marie-France Hivert*[,4,5,6], Luigi Bouchard**[,4,7,8] & Pierre-Étienne Jacques***[,1,4]

1. Département de Biologie, Université de Sherbrooke, Sherbrooke, Québec, JRK 2R1, Canada
2. Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, 94720-7360, USA
3. Département de Médecine, Université de Sherbrooke, Sherbrooke, Québec, J1K 2R1, Canada
4. Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Québec, J1H 5N4, Canada
5. Department of Population Medicine, Harvard Pilgrim Health Care Institute, Harvard Medical School, Boston, MA, 02115, USA
6. Diabetes Unit, Massachusetts General Hospital, Boston, MA, 02114, USA
7. Department of Biochemistry & Functional Genomics, Université de Sherbrooke, Sherbrooke, Québec, J1H 5N4, Canada
8. Department of Medical Biology, CIUSSS Saguenay-Lac-Saint-Jean, Hôpital de Chicoutimi, Saguenay, Québec, G7H 7K9, Canada

*Author for correspondence: MHivert@Partners.org

**Author for correspondence: Luigi.Bouchard@USherbrooke.ca

***Author for correspondence: Pierre-Etienne.Jacques@USherbrooke.ca

2.2.1 Abstract

Aim: The placenta goes through DNA methylation (DNAm) programming that is unique compared to all other fetal tissues. We aim to decipher some of the physiologic roles of the placenta by comparing its DNAm profile to another fetal tissue. **Materials & methods:** We performed a comparative analysis of genome-wide DNAm of 444 placentas paired with cord blood samples collected at birth. GO term analyses were conducted on the resulting differentially methylated regions. **Results:** Genomic regions upstream transcription start sites showing lower DNAm in the placenta were enriched with terms related to micro RNAs functions and genes encoding G protein-coupled receptors. **Conclusion:** These results highlight genomic regions that are differentially methylated in the placenta in contrast to fetal blood.

2.2.2 Keywords

DNA methylation, DNAm, placenta, cord blood, differentially methylated regions, DMR

2.2.3 Introduction

The placenta is responsible for maternal-fetal exchanges during pregnancy. It plays a crucial role in protecting the fetus from infection and actively participates in nutrient transfer to optimize fetal growth, in addition to influencing maternal physiology by secreting hormones [1]. Unsurprisingly, placental dysfunctions have been associated with various pregnancy complications such as pre-eclampsia and restricted fetal growth [2,3]. However, the interplay of maternal exposures, placental adaptive response, and other factors such as genetics is poorly understood and has only recently received attention [1,4]. Influenced by both genetics and environmental exposures, epigenetic marks such as DNA methylation (DNAm) [5] are central to our understanding of placental biology.

DNAm is a key mechanism of adaptive gene expression regulation, which is notably dynamic in embryonic tissues [6]. For this reason and because DNAm established early in life may be linked to long term health outcomes [7–9], the study of DNAm at birth from cord blood or placenta is becoming increasingly important. In itself, the unique placental DNAm profile,

which is globally hypomethylated relative to other tissues [10,11], has been previously described by different groups. However, only a handful of studies have performed a direct comparison of DNAm in the placenta with other fetal tissues. Those previous studies (including one from our group) generally had small sample sizes and/or focused on the similarities rather than the differences [12,13]. We hypothesized that analyzing placental DNAm in contrast with cord blood would be relevant to understand the distinctive placenta DNAm regulation landscape.

In the current study, we sought a better understanding of the fundamental biological roles of the placenta using a comparative genome-wide DNAm profiling approach. Based on the general principle that differences in DNAm can affect gene expression, we hypothesized that tissue-specific differentially methylated regions (tDMR) between the placenta and cord blood would highlight potential specific functions. To test this hypothesis, we identified tDMRs from Illumina EPIC array DNAm data of 444 paired umbilical cord blood and placenta samples.

2.2.4 Materials & methods

2.2.4.1 Cohort and sample collection

For this study, we used data from a subset of 444 participants from the Canadian cohort of *Genetics of Glucose regulation in Gestation and Growth* (Gen3G). Gen3G is a prospective pre-birth cohort recruited between 2010 and 2013 to investigate the pathophysiology of gestational diabetes and its impact on fetal development and children's health [14]. As previously described, we collected placenta tissue from the fetal side (1 cm^3, 5 cm from the umbilical cord insertion) and cord blood within 30 minutes of delivery [15]. The ethics review board of the *Centre Hospitalier Universitaire de Sherbrooke* approved all protocols, and written informed consent was obtained from all participants in agreement with the Declaration of Helsinki.

2.2.4.2 DNAm measurement and analysis

We performed DNAm measurement and preprocessing as previously described [15], independently in the placenta and cord blood. Briefly, we extracted DNA from cord blood and placenta samples, and we performed bisulfite conversion prior to DNAm quantification using the Infinium MethylationEPIC BeadChip (Illumina, San Diego, CA) that measures DNAm at over 850,000 CpGs across the genome. We performed quality control at the sample level to remove samples that had failed (n=8), or had mismatches on sex (n=1) or genotype (n=12), in addition to technical duplicates (n=20). Principal components of the control probes were used to process our data using functional normalization [16], and Regression on Correlated Probes (RCP) method [17] was used to adjust for probe-type bias before other quality controls at the probe level. We removed probes that had a non-significant detection *P*-value ($P > 0.05$) for ≥ 5% of the samples, and we corrected for batch/plate effects using the *ComBat* function from the *sva* package [18,19]. We also removed probes mapping to multiple locations, on sex chromosomes, flagged as non-CpG, or having a single nucleotide polymorphism (SNP) with a minor allele frequency ≥ 5% in their target sequence or at the single base extension. After applying these filters, placental and cord blood DNAm data shared 719,318 CpG sites; we used this subset for our analyses. We reported all our results using the hg19 assembly.

Unless specified otherwise, statistical analyses were performed using R version 3.5.1 [20], and most plots were generated using the package *ggplot2* version 3.1.0 [21]. We obtained gene annotations from *IlluminaHumanMethylationEPICanno.ilm10b4.hg19* version 0.6.0 [22]. We used M-values (logit-transformed β-values) in all analyses, for they are more appropriate to statistical analyses than β-values (e.g., reduced heteroscedasticity) [23].

2.2.4.2.1Multidimensional scaling

Multidimensional scaling is a method to visualize substructures in data of high dimensionality. It projects the data to a smaller dimensional space while attempting to preserve the distance between elements. Cord blood and placenta DNAm data from each participant were independently standardized (888 participant-tissue vectors), and pairwise Pearson's correlations

were calculated between all vectors. A distance matrix (1 - Pearson correlation coefficient) was then passed to the function *MDS* from the *sklearn* package [24], Python v3.6.0, to generate a visual representation. Each point was subsequently color-coded based on the tissue of origin.

2.2.4.2.2Cell type composition estimation

Reference-based methods for cell-type composition estimation are available for cord blood [25,26], but not for the placenta. To optimize the consistency of our cell type correction approach across tissues, we used ReFACTor [27], a reference-free approach, on both cord blood and the placenta. We used the seven first factors in each tissue to correct for cell-type heterogeneity in our analyses.

2.2.4.3 tDMR discovery and characterization

We used DMRcate version 1.18.0 [28] to identify tDMRs. The algorithm implemented in DMRcate allows us to analyze DNAm data measured with microarrays, identify differentially methylated regions with paired data (samples from the same individual), and is agnostic to genomic annotations. Briefly, DMRcate first computes a linear regression on each CpG site to test the effect of the variable of interest on DNAm (using *lmFit* and *eBayes* functions from *limma* [29]). Then, the moderated statistics from *limma* are squared and smoothed along the chromosomes using a Gaussian kernel. The resulting consecutive significant CpG sites are finally grouped to form regions if they are within 1 kb from each other. We used the default parameters for differential DNAm analysis in the functions *cpg.annotate* (linear regressions) and *dmrcate* (smoothing and grouping) from DMRcate. The linear regression model passed to *cpg.annotate* was:

$$Y \sim T + cbCT_{1,\dots,7} + pCT_{1,\dots,7} + I_{1,\dots,N} \qquad (2.1)$$

where N is the number of participants, Y is the vector of 2N DNAm values, T is the tissue variable (0 for cord blood, 1 for placenta), cbCT and pCT are the estimated cell-type composition for cord blood and the placenta, respectively, and I is the fixed effect for each participant. We also tested a model including sex of the baby, gestational age at delivery, as well

as maternal age, smoking status and early pregnancy body mass index as covariates. Because we observed very small differences in the significant CpG sites and filtered tDMRs we identified (~99% identical) after we added these covariates, we chose to report the results from the most parsimonious model. As an additional validation, we conducted all our analyses without including cell-type heterogeneity adjustment and found very similar results.

2.2.4.4 Gene ontology enrichment analysis

We performed GO term enrichment analyses with the function *gometh* from the R package *missMethyl* version 1.16.0 [30]. An advantage of this method is to consider the prior probability of gene enrichment caused by the unequal probe distribution among the genes due to array design. *Gometh* computes GO enrichments in the inputted list of significant probes using the "UCSC RefGene name" probe annotations from Illumina and the list of all probes included in the analysis.

To focus on potential genomic functionality, we categorized the tDMRs based on the tissue where the mean DNAm of the region was the lowest. We further stratified our analyses by the gene-relative position of the probes ("UCSC RefGene Group" annotation from Illumina) such as "TSS200", "TSS1500", or "first exon". Some probes had several gene-relative position annotations. However, sensitivity analyses performed by either removing these probes or assigning them to all relevant groups showed a negligible impact on the results. Therefore, we only report the results including all tDMRs probes. To reduce redundancy within the enriched terms, we retrieved the genes that contributed to the enrichment of each term, calculated the Jaccard distance between each list of genes, and used the R function *hclust* to cluster the lists hierarchically. The resulting dendrograms were then manually colored and annotated based on the descriptions of the GO terms.

The GO term enrichment results were consistent across multiple combinations of filtering thresholds (≥ 2, 3, 4, 5 or 10 CpG sites and ≥ 0.1, 0.2, 0.3, 0.4, or 0.5 absolute average β-value difference). We based our final threshold selection (> 4 CpG and absolute average β-value difference > 0.4) on the specificity of the enrichments in cb-tDMRs, which served as positive controls (Table S 2.1).

2.2.4.5 Gene expression analysis

To validate our tDMR results with gene expression data, we sought to find comparable expression datasets for cord blood and corresponding placenta samples. This proved challenging as there is a lack of publicly available gene expression datasets for the placenta and, to our knowledge, none from a study that investigated both tissues. To overcome this limitation, we used the uniformly reprocessed data from the recount2 portal [31] to identify a study conducted on cord blood-isolated CD34+CD45RA- cells (SRP027358, n=17) and one on healthy placenta samples (SRP068290, n=20). The data files were downloaded from the portal, and TPM normalized gene expression data was extracted using the function *getTPM* from the R/Bioconductor *recount* package [32]. In each dataset, we calculated the average TPM level across all samples for each gene after confirming the high correlation between samples (median Pearson correlation of ~97.4% and ~98.9% for SRP027358 and SRP068290, respectively). For graphical representation, we log10 transformed the average TPM values after adding a pseudocount (1e-5) to avoid zero values.

2.2.5 Results

2.2.5.1 Population characteristics

We analyzed DNAm data from 444 participants of the Gen3G cohort, a prospective Canadian pre-birth mother-child cohort [14]. As shown in Table 2.1, the women who participated in this study were on average 28 ± 4 years old at the beginning of pregnancy and had an early pregnancy body mass index of 25.5 ± 5.8. The minimal gestation age was 37 weeks (range 37-41), the third of the women were primigravid, and about half of the newborns were males (~53%). None of the pregnancies had major complications (e.g., pre-eclampsia; excluded for this study), and nearly all children were of European descent. At delivery, we collected umbilical cord blood and placenta samples from the fetal side, and DNAm was measured across the genome using the Infinium MethylationEPIC BeadChip. After quality control on the probes, 719,318 high-quality CpG probes across the genome were retained for further analyses.

Table 2.1 Characteristics of the Gen3G participants included in this study (n = 444)

	n (%)	Mean (SD)
Maternal age (years)	28.2 (4.3)	
Gravidity		
Primigravid (n)	146 (32.9)	
Smoking during pregnancy (n)		
No	400 (90.1)	
Yes	39 (8.8)	
Unknown	5 (1.1)	
Maternal BMI (kg/m^2)		25.5 (5.8)
Gestatioal age (weeks)		39.6 (1.0)
Male offspring (n)	235 (52.9)	

SD: Standard deviation; BMI: body mass index.

2.2.5.2 The placenta shows more partially methylated sites than cord blood and displays higher DNAm heterogeneity.

The DNAm value distribution was bimodal in both cord blood and the placenta, with one mode around 0.25% (low β-values) and the largest mode around 95% (high β-values) (Figure 2.1A). In the placenta, we also observed a density shift in the distribution from high DNAm levels to intermediate DNAm levels (Figure 2.1A and Figure S 2.1A). The differences in DNAm are also reflected in the unsupervised clustering of the data from both tissues, where we observed distinct clusters matching the tissue of origin (Figure 2.1B). Placenta DNAm values clustered less tightly, indicating greater DNAm heterogeneity in comparison to cord blood.

We plotted the distributions of DNAm values based on the gene-relative position of the sites. As expected, we observed in both tissues that sites within 200 bp upstream of the transcriptional start site (TSS) or in the first exon were more frequently unmethylated. In contrast, sites in the gene body or the 3' untranslated regions (UTR) tended to be more highly methylated (Figure S 2.2).

2.2.5.3 Most tDMRs have lower DNAm values in the placenta than in cord blood

We used DMRcate with the recommended parameters [28] to identify the tDMRs. Briefly, DNAm values at each site were regressed on the tissue variable while accounting for the estimated cell-type composition, and consecutive significant sites were grouped into tDMRs. Considering the differences observed in DNAm level distributions from both tissues (i.e., more intermediate values caused by a reduction of highly methylated sites in the placenta, Figure 2.1A), we expected a considerable proportion of the sites to be differentially methylated. The initial analysis identified 99,759 tDMRs (FDR < 0.05), composed of approximately 72% of all CpG sites analyzed (514,521 sites, Table S 2.2). The median length of the tDMRs was 780 bp (Figure S 2.3A), and most tDMRs (67,520 tDMRs, around 68%) included fewer than five probes (Figure 2.2A and Figure S 2.4A). For each tDMR, DMRcate reports the average β-value difference measured over the region. The median of these absolute average β-value differences was 0.11, and the larger regions tended to have smaller absolute average β-value differences (Figure 2.2A and Figure S 2.3B).

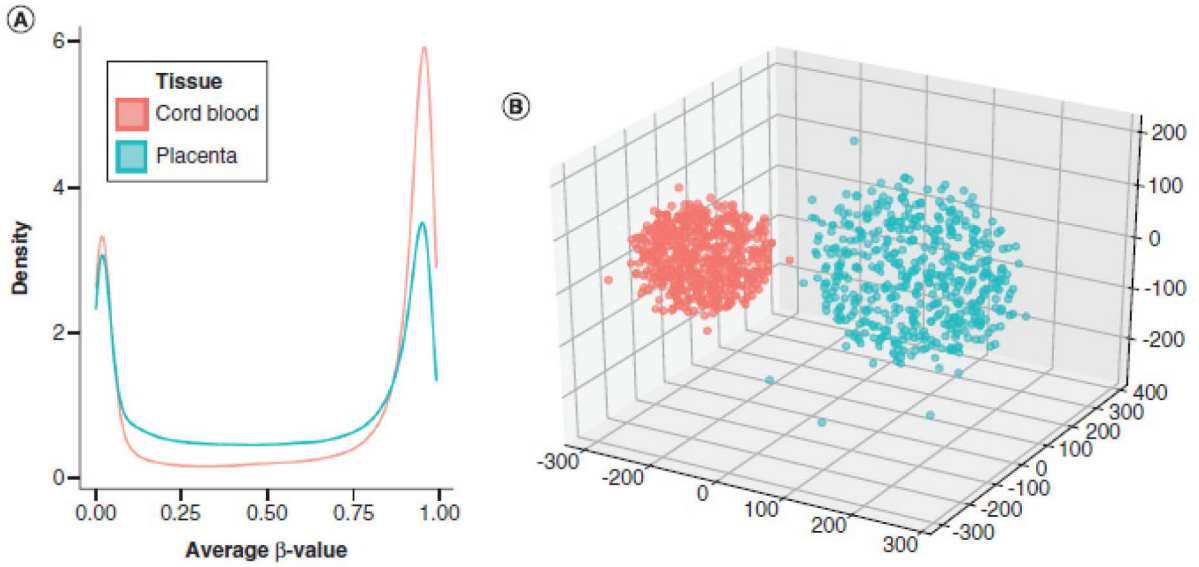

Figure 2.1 **Levels of DNAm are globally lower and more variable in the placenta than in cord blood.**

A) Distribution of the average β-values in cord blood (orange) and the placenta (cyan). The averages are calculated over the 444 participants for each CpG site and tissue. **B)** Multidimensional scaling (MDS) of the data, where each dot corresponds to a single participant in one of the two tissues (2 x 444 points).

We have focused our analyses on tDMRs with absolute average β-value differences larger than 0.40 that included more than three CpG sites, as they are more likely to be of biological relevance for a respective tissue. Using these thresholds, we identified a total of 4,912 tDMRs (Figure 2.2A). To ease the biological interpretability of our results, we categorized the tDMRs based on the tissue with the lowest DNAm values. Cord blood-tDMRs (cb-tDMRs, n=763 regions) were the regions with lower DNAm levels in cord blood than in the placenta (positive average β-value difference). Placenta-tDMR (p-tDMRs, n=4149 regions) were those with lower DNAm levels in the placenta than in cord blood (negative average β-value difference). Figure S 2.4 compares different characteristics of the filtered and unfiltered tDMRs.

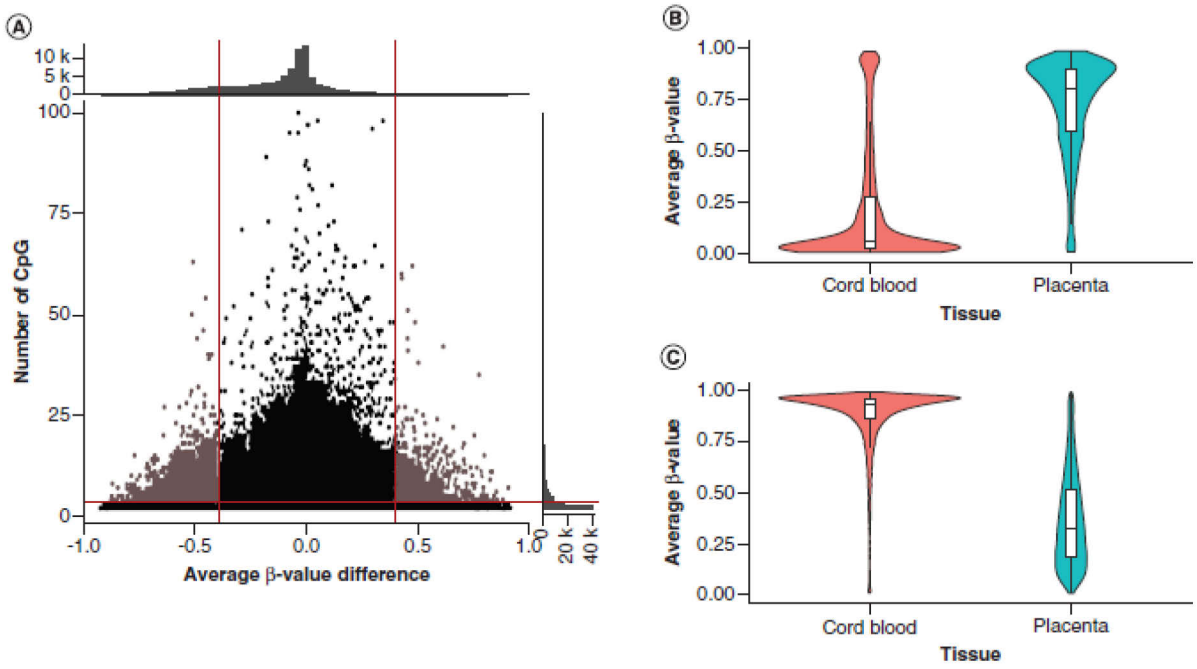

Figure 2.2 **Identification of cb-tDMRs and p-tDMRs.**

A) Relation between the number of CpG sites and the average β-value difference of a region. Red vertical lines indicate the average β-value difference thresholds (-0.4 and 0.4) used to select tDMRs, while the horizontal line indicates the CpG sites threshold (≥ 4). The regions selected for further analyses are colored in brown. 16 tDMRs with > 100 CpG sites are not represented in this panel. **B-C)** Average β-value distributions of the CpG sites included in filtered cb-tDMRs (B, 6,244 sites) or p-tDMRs (C, 25,868 sites) in cord blood (orange) and in the placenta (cyan). The averages are calculated in each tissue from the 444 participants.

Most sites in cb-tDMRs had high DNAm values in the placenta (> 0.7) and were mainly unmethylated in cord blood, with β-values close to zero (Figure 2.2B). In contrast, most sites in p-tDMRs had low (< 0.3) to intermediate (0.3-0.7) DNAm values in the placenta and high DNAm values in cord blood (Figure 2.2C and Figure S 2.1B). The DNAm profiles of both tDMR types showed specific characteristics, which we also observed when the sites were stratified by their gene-relative positions (Figure S 2.5).

2.2.5.4 cb-tDMRs and p-tDMRs are enriched with distinct Gene Ontology terms

More than two-third (3,351 out of 4,912) of the filtered tDMRs were annotated to at least one gene, as determined by probes annotations. We performed a Gene Ontology (GO) term enrichment analysis to assess the potential biological functions associated with these regions. Almost all identified GO categories were specific to either cb-tDMRs or p-tDMRs (Figure 2.3 and Table S 2.3). Lymphocyte activation and immune response were among the top GO categories in TSS1500- and TSS200-related sites in cb-tDMRs (Figure 2.3A-B). Figure 2.4A illustrates the well-known CD48 locus and the pattern of methylation we found in placenta and cord blood tissues. On the other hand, p-tDMRs showed enrichment in GO terms related to gene silencing by micro RNA, as well as sensory perception and G protein-coupled receptor activity (Figure 2.3C-D). Figure 2.4B illustrates our placenta-specific findings at a chr14 locus including *MEG3*, *MEG8*, and a large number of miRNAs clustered in the C14MC region. GO terms related to the plasma membrane were the only terms found in both cb-tDMRs and p-tDMRs.

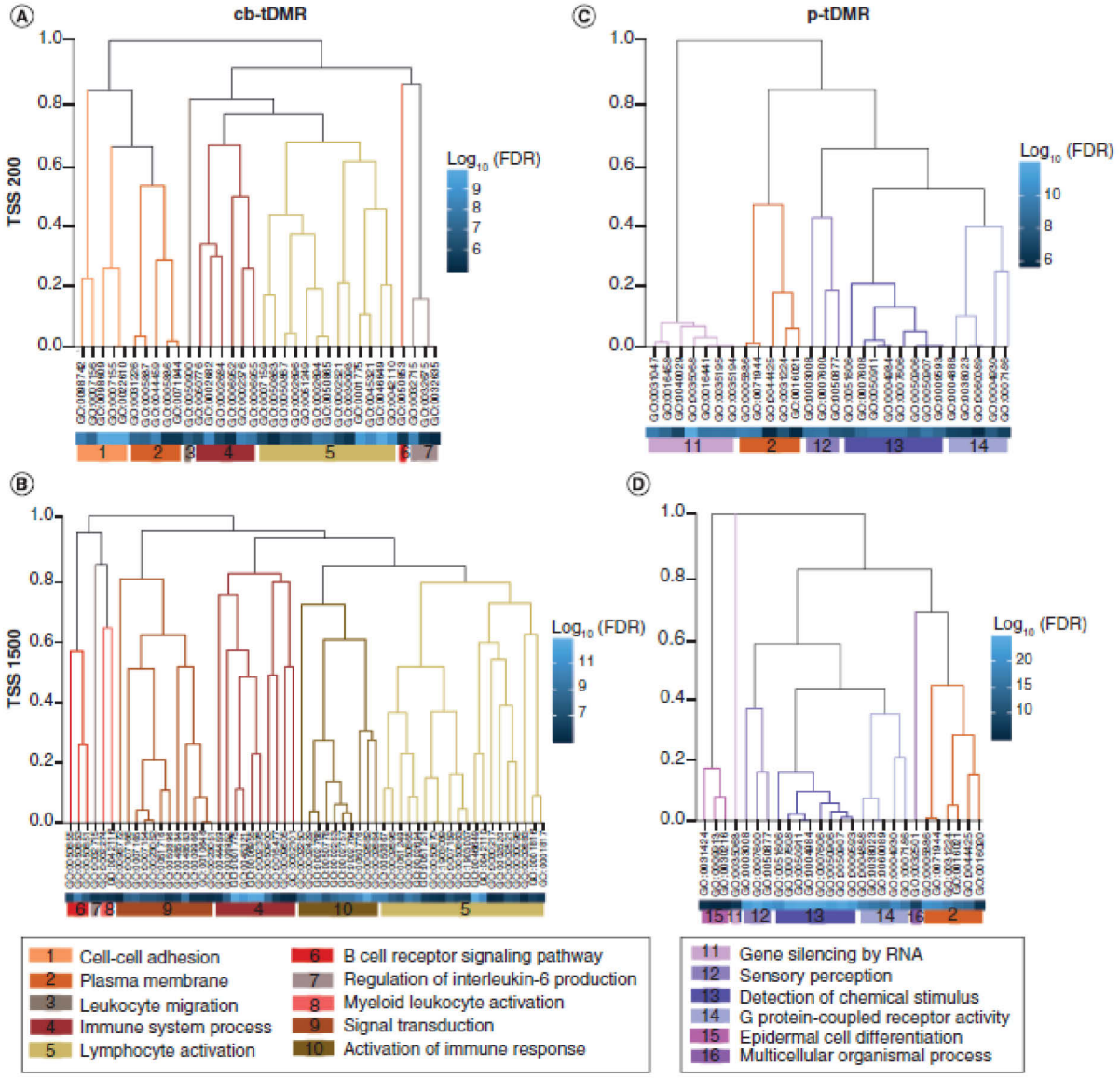

Figure 2.3 **Genes associated with cb-tDMRs and p-tDMRs differ in their functions.** Hierarchical clustering of GO terms based on their tDMR-specific contributing genes. **A-D)** GO terms associated with cb-tDMR sites (A-B) or p-tDMR sites (C-D) located < 200 bp (A, C) or between 200 bp and 1500 bp (B, D) upstream of their transcription start sites (TSS). Clusters were manually colored and annotated.

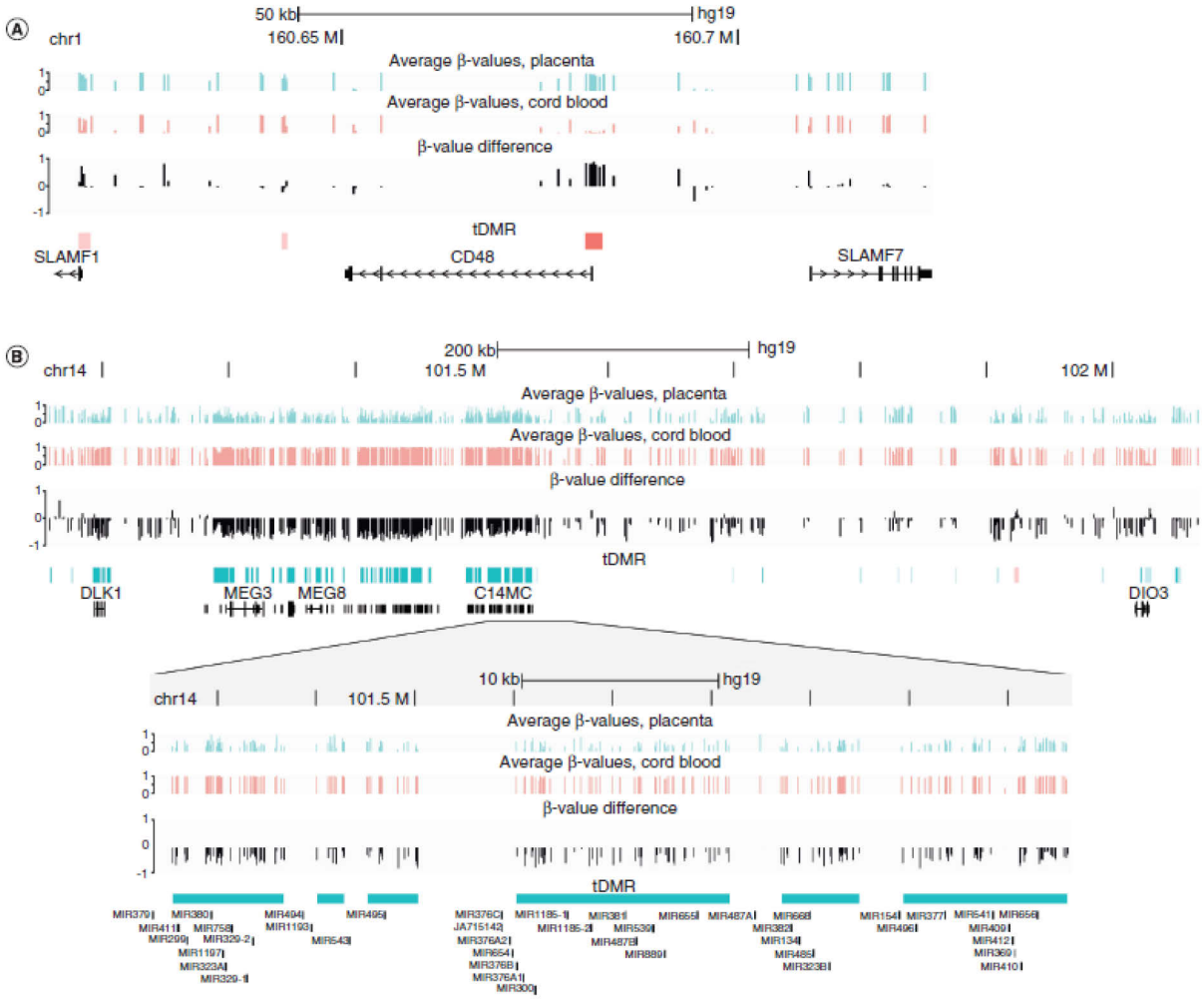

Figure 2.4 **cb-tDMR and p-tDMR examples.**

Examples of genomic regions including cb-tDMRs (A, CD48 protein-coding gene) and p-tDMRs (B, C14MC).

2.2.5.5 RNA expression in placenta and cord blood cells at cb-tDMRs and p-tDMRs associated genes

To support that our reported DNAm differences may also be linked to gene expression, we used two publicly available RNA-seq datasets from placenta and cord blood samples to assess the expression level of genes identified from p- or cb-tDMR-related sites within TSS200. We

compared expression levels of genes associated with the category of GO terms “Gene silencing by RNA” presented in Figure 2.3C (Figure 2.5A) or associated with the category “Immune system process” presented in Figure 2.3A (Figure 2.5B). We found that genes identified from p-tDMR in their promoter region (lower DNAm levels) had higher expression levels in placenta samples. In counterparts, genes with a cb-tDMR in their promoter region had higher expression levels in cord blood samples.

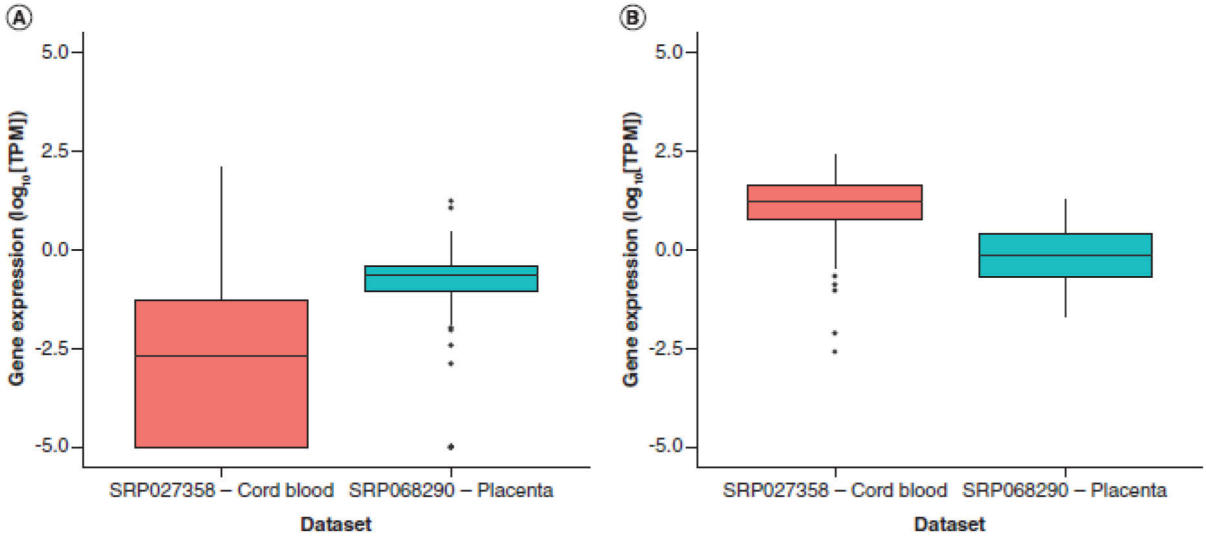

Figure 2.5 **Genes associated with cb-tDMRs and p-tDMRs are congruously expressed.**

Distribution of the log TPM-normalized expression level for the the 77 genes associated to Gene silencing by RNA (cluster 11 from Figure 2.3C) part of p-tDMRs (A), or the 81 genes associated to Immune system process (cluster 4 from Figure 3A) part of cb-tDMRs (B) in cord blood (orange) or placenta (cyan) samples. **A-B)** A pseudocount (1e-5) was added to all TMP values to avoid zero values prior to log transformation. cb: Cord blood; GO: Gene ontology; p: Placenta; tDMR: Tissue-specific differentially methylated region; TPM:Transcripts per million.

2.2.6 Discussion

In this study, we aimed to understand placental biology by comparing its DNAm profile to that of another fetal tissue, using cord blood cells as reference. We found more CpG sites with intermediate DNAm levels or high degrees of interindividual variability in the placenta than in cord blood. Our analyses revealed that most p-tDMRs were driven by genomic regions with low to intermediate DNAm values in the placenta and high DNAm values in cord blood. We focused our attention on filtered tDMRs located in TSS200 and TSS1500 (around promoter regions), since it is known that CpG sites at which DNAm is inversely correlated with expression are frequently found in these regions [33]. Pathway analyses of annotated genes at this subset of the p-tDMRs highlighted relevant functions in the placenta. Known features of the placenta methylome corroborated part of the regions we identified, as partially methylated domains [11] overlapped with ~63% of the p-tDMRs. Overall, our results are consistent with the literature on the placenta, e.g. the interindividual variability in placenta DNAm [6], the relative lower methylation of the placenta driving some of the p-tDMRs and the partial convergence of p-tDMRs and partially methylated domains.

Based on our results for cb-tDMRs, we found that GO term enrichments for tDMR within gene promoters are representative of tissue-specific functions. We showed that genes associated with cb-tDMRs were enriched for immune functions, which corresponds to an expected DNAm signature for leukocytes. Moreover, genes related to immune functions were more expressed in the cord blood dataset than they were in the placenta. This observation supports our hypothesis that tDMRs located in TSS200 and TSS1500 are likely to be informative on biological processes (i.e., reflect active promoters).

We have also captured distinctive differences in DNAm between the placenta and cord blood that improve our understanding of the placental biology. We found that p-tDMRs were enriched in GO terms related to miRNAs, including “micro-ribonucleoprotein complex” and “gene silencing by miRNA”. Genes from these GO categories were more expressed in the placenta than in cord blood, based on our analysis of two public datasets (Figure 2.5), which adds to the current literature on the importance of miRNAs in placental biology. The expression profile of

miRNAs in the placenta, for example clustered on chromosome 14 and 19, and their recognized or speculated roles were previously analyzed [34] and reviewed [35]. Interestingly, the miRNAs located in the chromosome 14 miRNA cluster (C14MC) had low DNAm levels in the placenta, and this region was clearly identified as p-tDMRs by our approach (Figure 2.4B). These miRNAs are known to be expressed in the placenta and are located within the *DLK1-DIO3* imprinted region, which also includes *MEG3* and *RTL1* genes [36]. There is some evidence that placental expression of these genes play a role in developmental programming, namely in the regulation of postnatal growth and energy balance [37]. As for the miRNAs in C14MC, it has been suggested that they might play key roles in developmental processes in placental mammals as they are only found in eutherians and as some are even associated with pregnancy pathologies [38]. However, their specific functions are still unclear [36,38].

Among our novel findings, GO terms related to sensory perception, the detection of chemical stimuli, and G protein-coupled receptors were also enriched in p-tDMRs. Many genes from the olfactory receptor (OR) family contributed to these enrichments. Whereas many OR genes are pseudogenized in humans, some of them are ectopically expressed [39]. Ectopically expressed ORs have diverse functions, e.g., in the lungs [40], kidneys [41], and other tissues as recently reviewed [42]. These findings in the human placenta may point to genes that may have a role in protecting the fetus against various adverse chemicals. Future functional studies will be needed to explore whether these ORs have true biological functions in the placenta.

Our study has limitations that we deem important to recognize. We are aware that gene expression regulation by DNAm is more complex than previously thought, in the placenta in particular [43]. Consequently, our GO enrichment results should not be directly extrapolated to gene expression. To counterbalance this limitation, we validated the difference in gene expression between cord blood and placenta samples in external datasets. While this analysis supports some of our findings, we are still limited by the fact that it could not be conducted on our samples. Our method is also restricted to differences detected at the tissue level; hence, differences driven by specific cell types are likely not reflected in our tDMRs. Moreover, we analyzed term placenta DNAm profiles, limiting our ability to interpret these results in the

context of the full length of pregnancy. Despite those limitations, our results proved robust to several sensibility analyses (e.g., filters, cell-type heterogeneity, covariates), and the enrichment of GO terms related to expected biological functions in cb-tDMRs supported our method. Our analyses were also conducted on a relatively large cohort of 444 pairs of placenta and cord blood samples, using the high-resolution 850k array for DNAm measurement.

In conclusion, we provided a unique analysis of tissue-level DNAm differences between placenta and cord blood, showing that a large portion of the human genome is differentially methylated between these two tissues. The tDMRs we identified were associated with tissue-specific biological functions, and genes annotated to these identified tDMRs were also differentially expressed in tissue in RNA analyses. Notably, we found that genes associated with p-tDMRs had functions related to miRNAs that are increasingly investigated for their roles during the pregnancy, and to the detection of chemical stimuli that open the door to novel investigations to better understand potential placental functions. These enrichments offer compelling avenues to explore the biological functions of the placenta and how they are regulated by its epigenome. Integrating multiple epigenomic layers in a single study, such as DNAm, histone modifications and gene expression, will eventually enable new insights into placental functions, related pregnancy adverse conditions, and health outcomes of the child.

2.2.7 Summary Points

- Comparative analysis of DNAm measured with Illumina MethylationEPIC BeadChip on 444 paired placenta and cord blood samples highlights differentially methylated regions associated with tissue-specific functions.
- Most p-tDMRs had low to intermediate DNAm levels in the placenta, which is consistent with the current knowledge on the placenta methylome.
- tDMRs within TSS200 and TSS1500 with lower DNAm in the placenta were associated with sensory perception, G protein-coupled receptors, and miRNA regulation, including the C14MC region.
- Genes associated with gene regulation by RNA that had a p-tDMR in their promoter were more expressed in the placenta than in cord blood in external datasets.

2.2.8 Author contributions

M Groleau conducted the analyses, interpreted findings and wrote the manuscript with the help of M-F Hivert, L Bouchard and P-E Jacques. F White and A Cardenas contributed to analyses. P Perron, M-F Hivert and L Bouchard contributed to the design and funding of the cohort, including for data/samples collection. M Groleau, M-F Hivert, L Bouchard and P-E Jacques contributed to the conception and design of the project. All authors revised the manuscript for important intellectual contributions and approved the final version of the manuscript.

2.2.9 Acknowledgments

The authors wish to thank research team members and all participants of the Gen3G cohort. They also wish to thank M Ahern and J Thompson for the English language revision of the manuscript. This research was partly enabled by the support provided by Calcul Québec and Compute Canada. M Groleau has a scholarship from the FRQNT, L Bouchard is a senior research scholar and P-E Jacques is a junior 2 research scholar from the FRQS. P-E Jacques, L Bouchard, M-F Hivert and P Perron are members of the FRQS-funded CRCHUS.

2.2.10 Financial & competing interests disclosure

This work was supported by American Diabetes Association accelerator award #1-15-AC x 10-26 (M-F Hivert), Fonds de Recherche du Québec en Santé (M-F Hivert #20697, L Bouchard, P-E Jacques #253444), Canadian Institute of Health Research #MOP 115071 (M-F Hivert), Diabète Québec (P Perron and L Bouchard) and Fonds de Recherche du Québec en Nature et Technologies #272263 (M Groleau). The authors declare no competing financial interests. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed. No writing assistance was utilized in the production of this manuscript.

2.2.11 Ethical conduct of research

The authors state that they have obtained appropriate institutional review board approval or have followed the principles outlined in the Declaration of Helsinki for all human or animal experimental investigations. In addition, for investigations involving human subjects, informed consent has been obtained from the participants involved.

2.2.12 Data sharing statement

We provide the complete list of the tDMRs we identified and the CpG included in each tDMRs in the supplementary material accompanying this article. Other data is available upon request, under ethical approbation.

2.2.13 Open access

2.2.14 References

Papers of special note have been highlighted as: • of interest; •• of considerable interest.

1. Maltepe E, Fisher SJ. Placenta: The Forgotten Organ. *Annu. Rev. Cell Dev. Biol.* [Internet]. 31(1), 523–552 (2015). Available from: www.annualreviews.org.

2. Manokhina I, Del Gobbo GF, Konwar C, Wilson SL, Robinson WP. Review: placental biomarkers for assessing fetal health. *Hum. Mol. Genet.* [Internet]. 26(R2), R237–R245 (2017). Available from: http://academic.oup.com/hmg/article/26/R2/R237/3865112/Review-placental-biomarkers-for-assessing-fetal.

3. Cuffe JSM, Holland O, Salomon C, Rice GE, Perkins A V. Review: Placental derived biomarkers of pregnancy disorders. *Placenta* [Internet]. 54, 104–110 (2017). Available from: https://www.sciencedirect.com/science/article/pii/S0143400417301212?via%3Dihub.

4. Perez-Garcia V, Fineberg E, Wilson R, *et al.* Placentation defects are highly prevalent in embryonic lethal mouse mutants. *Nature* [Internet]. 555(7697), 463–468 (2018). Available from: http://www.nature.com/articles/nature26002.

5. Vaiman D. Genes, epigenetics and miRNA regulation in the placenta. *Placenta* [Internet]. 52, 127–133 (2017). Available from: https://www.sciencedirect.com/science/article/pii/S0143400416306804?via%3Dihub#bib19.

6. Bianco-Miotto T, Mayne BT, Buckberry S, Breen J, Rodriguez Lopez CM, Roberts CT. Recent progress towards understanding the role of DNA methylation in human placental development. *Reproduction* [Internet]. 152(1), R23-30 (2016). Available from: http://www.ncbi.nlm.nih.gov/pubmed/27026712.

7. Cecil CAM, Walton E, Jaffee SR, *et al.* Neonatal DNA methylation and early-onset conduct problems: A genome-wide, prospective study. *Dev. Psychopathol.* [Internet]. 30(2), 383–397 (2018). Available from: https://www.cambridge.org/core/product/identifier/S095457941700092X/type/journal_article.

8. Seow WJ, Ngo CS, Pan H, *et al.* In-utero epigenetic factors are associated with early-onset myopia in young children. *PLoS One*. 14(5) (2019).

9. Cardenas A, Lutz SM, Everson TM, Perron P, Bouchard L, Hivert MF. Mediation by Placental DNA Methylation of the Association of Prenatal Maternal Smoking and Birth Weight. *Am. J. Epidemiol.* 188(11), 1878–1886 (2019).

10. Robinson WP, Price EM. The human placental methylome. *Cold Spring Harb. Perspect. Med.* 5(5), 1–15 (2015).

11. Schroeder DI, Blair JD, Lott P, *et al.* The human placenta methylome. *Proc. Natl. Acad. Sci.* [Internet]. 110(15), 6037–6042 (2013). Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1215145110.

    •• Comprehensive analysis of the placental methylome and its characteristic features, such as partially methylated domains.

12. Ma B, Allard C, Bouchard L, *et al.* Locus-specific DNA methylation prediction in cord blood and placenta. *Epigenetics*. 14(4), 405–420 (2019).

13. De Carli MM, Baccarelli AA, Trevisi L, *et al.* Epigenome-wide cross-tissue predictive modeling and comparison of cord blood and placental methylation in a birth cohort. *Epigenomics* [Internet]. 9(3), 231–240 (2017). Available from: http://www.futuremedicine.com/doi/10.2217/epi-2016-0109.

14. Guillemette L, Allard C, Lacroix M, *et al.* Genetics of Glucose regulation in Gestation and Growth (Gen3G): A prospective prebirth cohort of mother-child pairs in Sherbrooke, Canada. *BMJ Open*. 6(2), 1–14 (2016).

15. Cardenas A, Gagné-Ouellet V, Allard C, *et al.* Placental DNA Methylation Adaptation to Maternal Glycemic Response in Pregnancy. *Diabetes*. 67(8), 1673–1683 (2018).

16. Fortin J-P, Labbe A, Lemire M, *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* [Internet]. 15 (2014). Available from: http://genomebiology.com/2014/15/11/503.

17. Niu L, Xu Z, Taylor JA. RCP: a novel probe design bias correction method for Illumina Methylation BeadChip. *Bioinformatics* [Internet]. 32(17), 2659–63 (2016). Available from: http://www.ncbi.nlm.nih.gov/pubmed/27153672.

18. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* [Internet]. 28(6), 882–883 (2012). Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts034.

19. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* [Internet]. 8(1), 118–127 (2007). Available from: https://academic.oup.com/biostatistics/article/8/1/118/252073.

20. R Core Team. R: A language and environment for statistical computing. [Internet]. (2018). Available from: https://www.r-project.org/.

21. Wickham H. ggplot2 : Elegant Graphics for Data Analysis [Internet]. Springer International Publishing, Cham. Available from: https://ggplot2.tidyverse.org.

22. Hansen KD. IlluminaHumanMethylationEPICanno.ilm10b4.hg19: Annotation for Illumina's EPIC methylation arrays [Internet]. (2017). Available from: https://bitbucket.com/kasperdanielhansen/Illumina_EPIC.

23. Du P, Zhang X, Huang CC, *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* [Internet]. 11(1), 587 (2010). Available from: http://www.biomedcentral.com/1471-2105/11/587.

24. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* [Internet]. 12, 2825–2830 (2011). Available from: http://scikit-learn.sourceforge.net.

25. Lin X, Tan JYL, Teh AL, *et al.* Cell type-specific DNA methylation in neonatal cord tissue and cord blood: a 850K-reference panel and comparison of cell types. *Epigenetics*. 13(9) (2018).

26. Cardenas A, Allard C, Doyon M, *et al.* Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics* [Internet]. 11(11), 773–779 (2016). Available from: http://dx.doi.org/10.1080/15592294.2016.1233091.

27. Rahmani E, Zaitlen N, Baran Y, *et al.* Sparse PCA corrects for cell type heterogeneity in epigenome-wide association studies. *Nat. Methods* [Internet]. 13(5), 443–445 (2016). Available from: http://www.nature.com/articles/nmeth.3809.

28. Peters TJ, Buckley MJ, Statham AL, *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics Chromatin* [Internet]. 8(1), 6 (2015). Available from: http://www.epigeneticsandchromatin.com/content/8/1/6.

    •• Description of the main analytical tool (DMRcate) used in this study.

29. Ritchie ME, Phipson B, Wu D, *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* [Internet]. 43(7), e47–e47 (2015). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/pdf/gkv007.pdf.

30. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* [Internet]. 32(2), btv560 (2015). Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv560.

31. Collado-Torres L, Nellore A, Kammers K, *et al.* Reproducible RNA-seq analysis using recount2. *Nat. Biotechnol.* [Internet]. 35(4), 319–321 (2017). Available from: https://jhubiostatistics.shinyapps.io/recount/.

32. Collado-Torres L, Nellore A, Jaffe AE. recount workflow: Accessing over 70,000 human RNA-seq samples with Bioconductor. *F1000Research* [Internet]. 6, 1558 (2017). Available from: https://doi.org/10.12688/f1000research.12223.1.

33. Delahaye F, Do C, Kong Y, *et al.* Genetic variants influence on the placenta regulatory landscape. *PLOS Genet.* [Internet]. 14(11), e1007785 (2018). Available from: https://dx.plos.org/10.1371/journal.pgen.1007785.

34. Morales-Prieto DM, Chaiwangyen W, Ospina-Prieto S, *et al.* MicroRNA expression profiles of trophoblastic cells. *Placenta* [Internet]. 33(9), 725–734 (2012). Available from: https://www.sciencedirect.com/science/article/pii/S0143400412002081?via%3Dihub.

35. Poirier C, Desgagné V, Guérin R, Bouchard L. MicroRNAs in Pregnancy and Gestational Diabetes Mellitus: Emerging Role in Maternal Metabolic Regulation. *Curr. Diab. Rep.* [Internet]. 17(5), 35 (2017). Available from: http://link.springer.com/10.1007/s11892-017-0856-5.

36. Malnou EC, Umlauf D, Mouysset M, Cavaillé J. Imprinted microRNA gene clusters in the evolution, development, and functions of mammalian placenta. Front. Genet.10(JAN) (2019).

    • Review of the essential roles of miRNAs in mammalian placenta and development.

37. Prats-Puig A, Carreras-Badosa G, Bassols J, *et al.* The placental imprinted DLK1-DIO3 domain: a new link to prenatal and postnatal growth in humans. *Am. J. Obstet. Gynecol.* [Internet]. 217(3), 350.e1-350.e13 (2017). Available from: https://www.sciencedirect.com/science/article/pii/S0002937817306075?via%3Dihub.

38. Morales-Prieto DM, Ospina-prieto S, Chaiwangyen W, Schoenleben M, Markert UR. Pregnancy-associated miRNA-clusters. *J. Reprod. Immunol.* [Internet]. 97(1), 51–61 (2013). Available from: http://dx.doi.org/10.1016/j.jri.2012.11.001.

39. Flegel C, Manteniotis S, Osthold S, Hatt H, Gisselmann G. Expression Profile of Ectopic Olfactory Receptors Determined by Deep Sequencing. *PLoS One* [Internet]. 8(2), e55368 (2013). Available from: http://dx.plos.org/10.1371/journal.pone.0055368.

40. An SS, Liggett SB. Taste and smell GPCRs in the lung: Evidence for a previously unrecognized widespread chemosensory system. Cell. Signal.41, 82–88 (2018).

41. Shepard BD, Pluznick JL. How does your kidney smell? Emerging roles for olfactory receptors in renal function. Pediatr. Nephrol.31(5), 715–723 (2016).

42. Maßberg D, Hatt H. Human olfactory receptors: Novel cellular functions outside of the nose. Physiol. Rev.98(3), 1739–1763 (2018).

43. Lim YC, Li J, Ni Y, *et al.* A complex association between DNA methylation and gene expression in human placenta at first and third trimesters. *PLoS One*. 12(7), 1–15 (2017).

2.2.15 Supplemental material

2.2.15.1 Supplemental figures

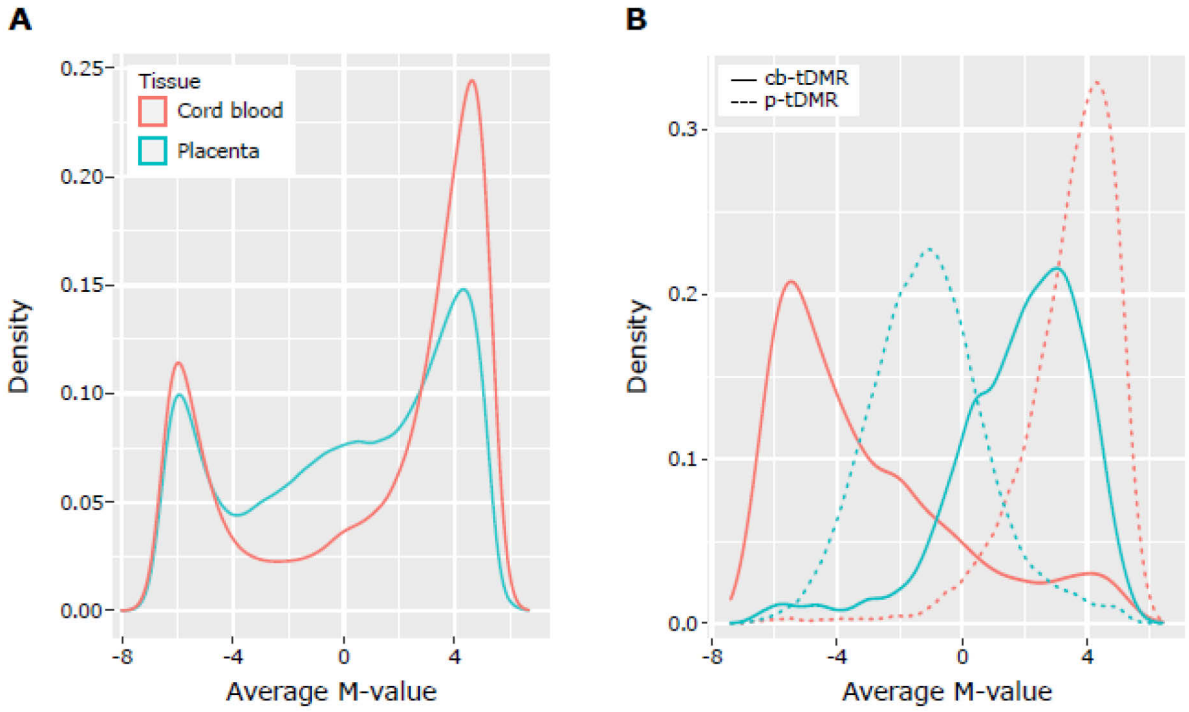

Figure S 2.1 **Distribution of the average M-values in cord blood and the placenta for every CpG site analyzed (719,318, A), and for the CpG sites included in the filtered cb-tDMRs (6,244) and p-tDMRs (25,868) (B).**

Supporting Figure 2.1A and 2.2B.

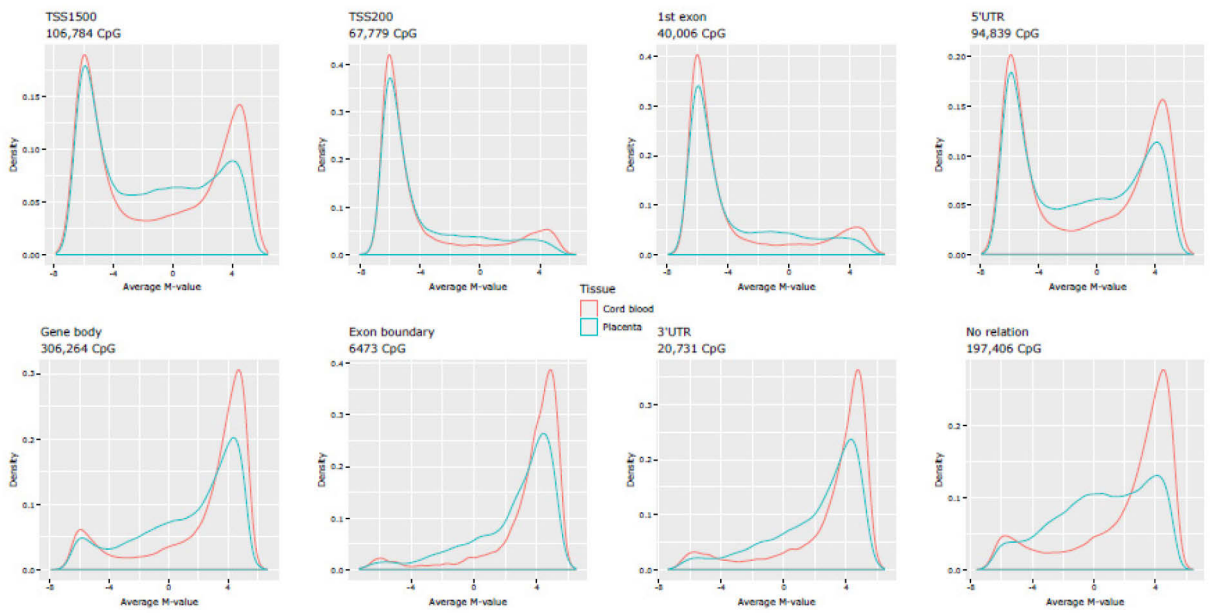

Figure S 2.2 **Distributions of average M-values in cord blood and the placenta based on the gene-relative position of all the CpG sites.**

The categories are as annotated by Illumina. TSS1500: 200-1500 bases upstream of the TSS; TSS200: 0-200 bases upstream of the TSS; 5'UTR: within the 5' untranslated region, between the TSS and the ATG start site; Gene body: between the ATG and stop codon, irrespective of the presence of introns, exons, TSS, or promoters; 3'UTR: between the stop codon and the poly(A) site. Note that definitions for "First exon" and "Exon boundary" are not explicitly in the documentation, and they are the only categories overlapping with others. CpG sites associated with multiple categories are included in each corresponding graph.

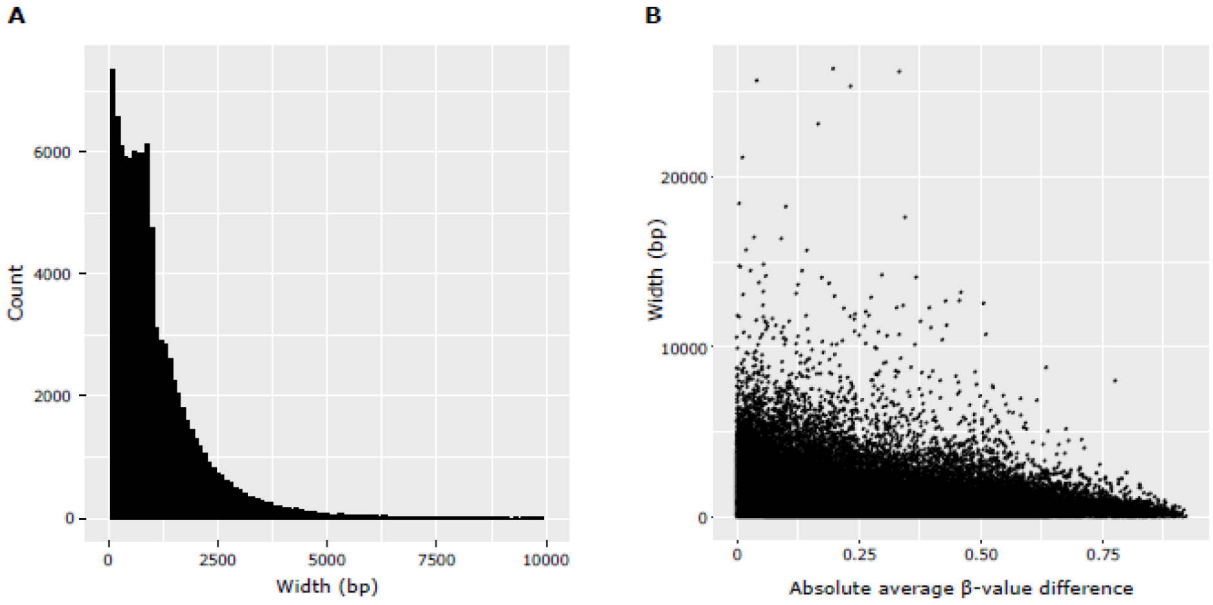

Figure S 2.3 **Descriptive characteristics for the whole set of 99,759 unfiltered tDMRs.** **A)** Distribution of tDMR lengths. 90 tDMRs longer than 10,000 bp are not represented in this figure. **B)** The relation between the length and the absolute average β-value difference of a tDMR.

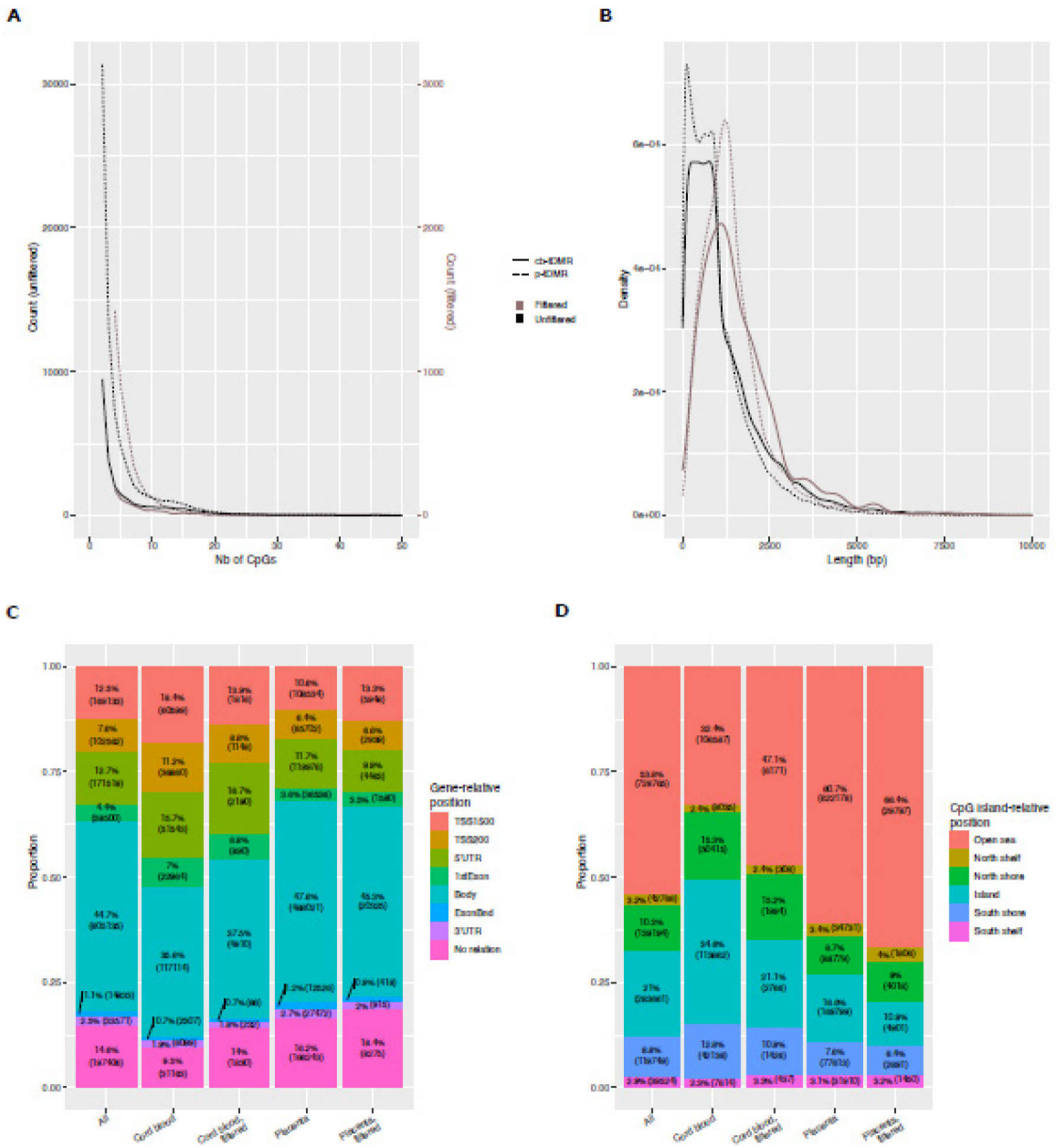

Figure S 2.4 **Descriptive characteristics of filtered and unfiltered tDMRs.**

A) Number of tDMRs based on the number of CpG sites they include. 70 cb-tDMR (2 "Filtered") and 44 p-tDMRs (2 "Filtered") encompass more than 50 CpG sites and are not represented in this figure. **B)** Distribution of tDMRs based on their lengths. 59 cb-tDMRs (3 "Filtered") and 31 p-tDMRs (4 "Filtered") are longer than 10,000 bp and are not represented in this figure. **C-**

D) Proportions of gene-relative (C) and CpG island-relative (D) positions of sites included in (un)filtered tDMRs compared to all sites in the analysis. Sites with multiple gene-relative position annotations were classified according to their most frequent annotation.

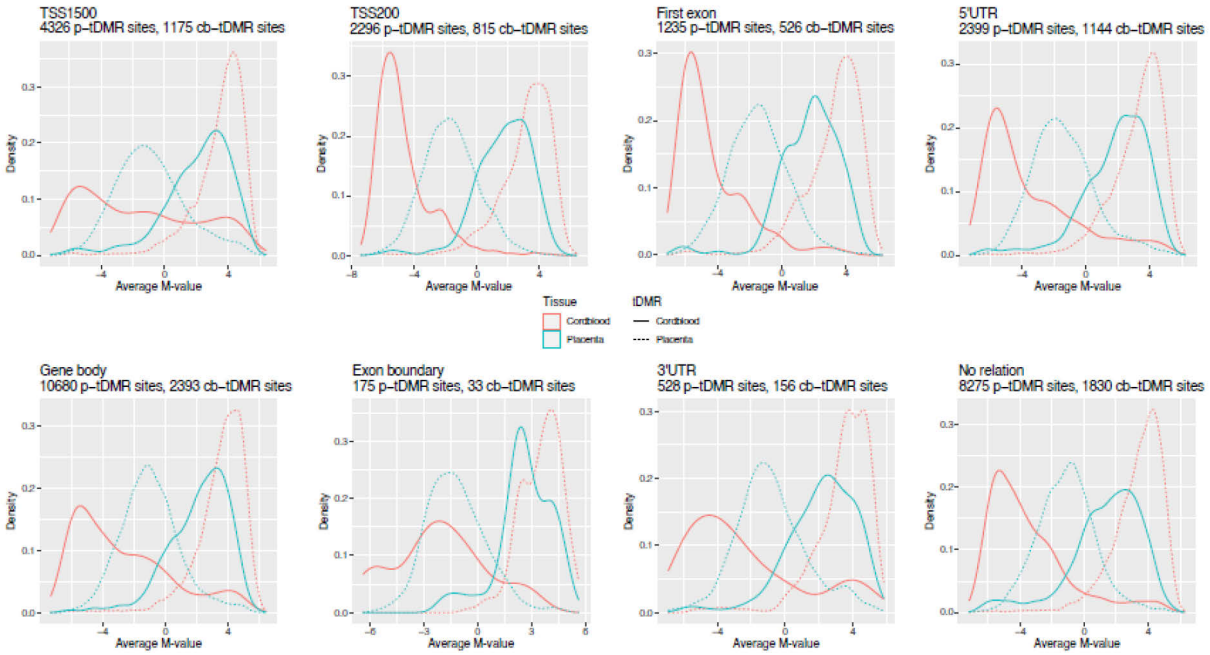

Figure S 2.5 **Distributions of average M-values in cord blood and the placenta of sites included in cb- or p-tDMRs, categorized by their gene-relative positions.** As in Figure S 2.2.

2.2.15.2 Supplemental tables

Tables S 2.1-2.3 (supplementary tables 1-3 in the published article) can be found online at : www.futuremedicine.com/doi/suppl/10.2217/epi-2020-0271.

Table S 2.1 **GO term enrichments at different combinations of CpG number and average β-value difference thresholds.**

Table S 2.2 **Identified tDMRs and the CpGs they include.**

Table S 2.3 **GO term enrichments of the filtered tDMRs at different gene-relative positions.**

CHAPITRE 3
CARTOGRAPHIE PANGÉNOMIQUE DES LOCI QUANTITATIFS DE MÉTHYLATION DE L'ADN DANS LE PLACENTA

3.1 Introduction

Comme décrit à la section 1.6.3, la visée de cette seconde étape de mon projet de recherche de maîtrise était de cartographier les mQTL dans le placenta puis d'en faire une analyse descriptive avant de finalement les utiliser pour revisiter des analyses déjà publiées avec les données de la cohorte Gen3G.

3.2 Matériel et méthodes

3.2.1 Cohorte

Les données utilisées pour ce projet proviennent elles aussi de la cohorte Gen3G, telle que décrite au Chapitre 2. Le sous-ensemble de participants diffère cependant dû à la disponibilité des données requises. Un sous-ensemble de 524 participantes a été sélectionné pour la disponibilité des données de mADN, et un de 434 pour celles des données génétiques. Le Tableau 3.1 présente les caractéristiques des participantes dans l'union (N = 533) ou l'intersection (N = 425) de ces deux sous-ensembles.

3.2.2 Données de mADN

Les données de mADN ont été produites avec la puce EPIC de Illumina à partir de biopsies de placenta du côté fœtal, tel que décrit dans une publication antérieure (Cardenas *et al*., 2018) ainsi qu'au Chapitre 2. Pour ce projet, les étapes de contrôle de qualité et de pré-traitement des données des 566 échantillons ont été harmonisées avec celles du consortium *Pregnancy And Childhood Epigenetics* (PACE) (Felix *et al*., 2018), dont fait partie la cohorte Gen3G. Les étapes

Tableau 3.1 **Caractéristiques des participantes de Gen3G et de leurs enfants, dans l'union et l'intersection des sous-ensembles de données génétiques et épigénétiques.**

	Union (N = 533)		Intersection (N = 425)	
	Nb (%)	**Moyenne (é-t)**	**Nb (%)**	**Moyenne (é-t)**
Âge mère (an)		28,6 (4,3)		28,8 (4,2)
Ethnicité				
Européen	527 (98,9)		422 (99,3)	
Fumeuse				
Oui	47 (8,8)		35 (8,2)	
Non	481 (90,2)		385 (90,6)	
Inconnu	5 (0,9)		5 (1,2)	
IMC (kg/m^2)		24,8 (5,7)		24,7 (5,7)
Nb de grossesses				
Primigravide	183 (34,3)		138 (32,5)	
Sexe enfant				
Fille	250 (46,9)		202 (47,5)	
Garçon	281 (52,7)		223 (52,5)	
Manquant	2 (0,4)		0 (0)	
Âge gestationnel (semaines)		39,5 (1,2)		39,5 (1,0)

é-t : écart-type; IMC : indice de masse corporelle.

sont détaillées sur leur site web (Epi-centered Research, n.d.) et ont toutes été effectuées sur les valeurs bêta. En bref, les échantillons n'atteignant pas les standards de qualité d'Illumina (N = 7), présentant une discordance entre leur sexe estimé et celui connu (N = 6) ou ayant potentiellement été contaminés avec du matériel maternel (N = 2) ont été retirés. Les échantillons ayant une discordance de signature génétique, telle que déterminée avec les

contrôles de génotypage intégrés à la puce, ont aussi été retirés (N = 7) (Heiss et Just, 2018). Parmi les duplicatas techniques, les échantillons avec les plus faibles ratios médians de signal de méthylation ont été retirés (N = 15). Le contrôle de qualité des sondes a permis de retirer un échantillon ayant trop de sondes n'atteignant pas le seuil de qualité requis. Finalement, les échantillons présentant de faibles taux de syncytiotrophoblastes après estimation de leur composition cellulaire par la méthode *planet* (Yuan *et al*., 2021) ont été retirés (N = 4).

Ensuite, les sondes ont été exclues de l'ensemble de données si elles n'avaient pas une lecture de qualité suffisante dans 95% des échantillons (N = 84 316). L'ensemble de 524 échantillons et de 781 543 sondes a ensuite été traité pour corriger le bruit de fond, la variation due aux manipulations ainsi que les biais introduits par le type de sonde en combinant les approches de *minfi* (Fortin *et al*., 2017), *Noob* (Triche *et al*., 2013), *beta-mixture quantile* (Teschendorff *et al*., 2013) et de normalisation fonctionnelle (Fortin *et al*., 2014). *ComBat* a également été utilisé pour corriger pour l'effet de lot (Johnson *et al*., 2007). La valeur des sondes pour lesquelles la p-value de détection était > 5% a été considérée comme manquante, puis le 1% des valeurs extrêmes a été winsorisé. Cette étape consiste à remplacer les valeurs dépassant le percentile seuil (0,5% et 95,5% pour un seuil de 1%) par la valeur du percentile, ce qui « écrase »- les valeurs extrêmes sur le reste de la distribution.

Après l'application du pipeline de PACE, d'autres étapes de filtres ont été ajoutées afin de répondre à nos critères d'analyse. Ces étapes sont également décrites dans le Chapitre 2. Les sondes sur les chromosomes sexuels, ne mesurant pas un site CpG, ayant un SNP avec une fréquence allélique >= 5% dans leur séquence cible ou au nucléotide d'extension ou s'hybridant à plusieurs endroits dans le génome ont été exclues. De plus, les CpG présentant une faible variabilité dans la mADN mesurée à travers les individus ont été retirées. Cette étape permet de limiter le nombre de régressions calculées et ne limite pas notre capacité à identifier des mQTL, comme on cherche des CpG dont les niveaux de mADN varient selon le génotype. Le seuil de variance limite qui a été utilisé est de 1×10^{-5}, valeur qui élimine approximativement le 5% le moins variable des CpG. Au final, l'ensemble de données contenait 681 795 CpG et 524 individus.

3.2.3 Données génétiques

Les données génétiques ont été obtenues par séquençage de génome entier avec la technologie d'Illumina. Brièvement, l'alignement a été effectué sur l'assemblage GRCh37 et la suite GATK 3.8 a été utilisée avec quelques modifications par rapport aux recommandations d'utilisation (DePristo *et al*., 2011), notamment en omettant l'étape de recalibration des bases. Les génotypes ont été assignés pour chaque position génomique pour chaque individu par génotypage groupé (*joint genotyping*) (Poplin *et al*., 2017). Pour nos analyses, des filtres et contrôles supplémentaires ont été appliqués sur les individus ainsi que sur les positions. Pour chaque position, pour chaque individu, le génotype a été considéré comme manquant si moins de 10 lectures y ont été alignées. Ensuite, toutes les positions où > 20% des données sont manquantes, dévient de l'équilibre de Hardy-Weinberg ($P < 10\text{-}6$ du test exact) ou comptant moins de 10 occurrences de l'allèle alternatif dans notre cohorte ont été retirées avec PLINK 1.9 (Chang *et al*., 2015). Les duplicatas techniques ainsi que les individus présentant un taux de données manquantes plus grand que 20% ou des signes de contamination (grande hétérozygocité) ont été retirés. Suite à l'estimation de la parentalité et de l'ethnicité (section 1.3.4), 22 individus apparentés à un autre participant de la cohorte ainsi que 10 individus d'ascendance génétique trop distincte ont été retirés. Les méthodes d'estimation de la parentalité et de l'ethnicité sont décrites ci-bas. L'ensemble final comprend 405 individus et ~8,8M positions.

3.2.4 Évaluation de l'ethnicité, de la parentalité et de la structure populationnelle

Dans ce type d'analyse génétique sur des cohortes humaines, il est attendu qu'il existe de la parentalité inconnue des participants ainsi que des inexactitudes dans l'ascendance auto-rapportée. La suite d'outils R *Genesis* (Gogarten *et al*., 2019), disponible sur Bioconductor, permet d'estimer de manière robuste la parentalité ainsi que l'ascendance génétique d'une cohorte. Plus spécifiquement, *PC-AiR* (Conomos *et al*., 2015) permet d'estimer l'ascendance génétique en présence de parentalité et *PC-Relate* (Conomos *et al*., 2016), la parentalité en présence de structure populationnelle. L'outil *PC-AiR* a donc été utilisé pour estimer l'ascendance génétique de la population en croisant les données de la cohorte Gen3G avec celles

de la phase 1 du *1000 Genomes Projects* (Altshuler *et al*., 2012). À cette fin, un sous-ensemble de SNP indépendants (valeur de corrélation paire à paire dans une fenêtre de 1 Mb < 0,32), présentant < 1% de données manquantes et ayant une fréquence allélique > 5% dans l'ensemble groupé des données de Gen3G et de 1000G a été utilisé. Les individus s'écartant de plus de 5 écarts-types de la moyenne des populations TSI (Toscans d'Italie) et CEU (résidents de l'Utah d'ascendance nord et est-européenne) pour les composantes principales 1 ou 2 ont été considérés comme trop divergents (N = 10). Ensuite, *PC-Relate* a été utilisé pour estimer la parentalité, en utilisant cette fois-ci seulement les individus de Gen3G ayant passé le filtre sur l'ascendance génétique. Le sous-ensemble de SNP était composé de SNP indépendants (valeur de corrélation paire à paire dans une fenêtre de 1 Mb < 0.32), présentant < 1% de données manquantes, ayant une fréquence allélique de > 5% et ne déviant de l'équilibre de Hardy-Weinberg qu'au maximum avec une valeur $p > 10^{-4}$ (test exact). Le seuil de coefficient de relation de parentalité (*coefficient of relationship*) utilisé pour sélectionner des individus non apparentés était de moins de 2^-5,5 (plus éloigné qu'un second cousin). Un individu de chacune de ces paires, celui présentant les données de séquençage de plus faible qualité, a été retiré des analyses pour obtenir un ensemble d'individus non apparentés (N = 22). Finalement, avec le même ensemble de SNP que pour l'estimation de la parentalité, *PC-AiR* a été utilisé pour calculer les composantes principales de la structure populationnelle de Gen3G, laquelle s'avérait être assez homogène.

3.2.5 Méthode statistique pour l'identification de mQTL

Il existe plusieurs outils bio-informatiques permettant d'identifier des QTL, incluant les mQTL, comme revus dans (Ye *et al*., 2020). Pour ce projet, c'est l'outil *TensorQTL* (Taylor-Weiner *et al*., 2019) qui a été retenu, une version plus rapide (car optimisée pour être exécutée sur des unités de calcul graphique) de *FastQTL* (Ongen *et al*., 2016). En plus de son efficacité, il a l'avantage d'être utilisé dans le pipeline bio-informatique de GTEx, sur lequel une partie de la méthodologie de ce projet est basée (Aguet *et al*., 2020). Comme les associations recherchées sont proximales, une distance d'au plus 500 Kb entre un site CpG et un SNP était admise pour calculer une association, ce qui est cohérent avec le standard qui varie généralement entre 100 Kb et 1 Mb dans la littérature (Gaunt *et al*., 2016). Parmi les différentes méthodes offertes par

TensorQTL pour calculer des QTL proximaux, nous avons utilisé la fonction *cis.map_cis*, laquelle produit une seule statistique par phénotype (en l'occurrence, par CpG). Pour ce faire, une régression est calculée pour chaque paire de SNP-CpG, puis une distribution bêta est générée par permutation pour chaque CpG afin d'identifier, s'il y a lieu, le SNP pour lequel l'association est la plus forte et significative. Les valeurs *p* obtenues par cette dernière étape sont ensuite ajustées pour prendre en considération les tests multiples par la méthode de (Storey et Tibshirani, 2003) afin d'obtenir des valeurs *q*. Le seuil de significativité des valeurs *q* a été établi à 10%. Le modèle de régression pour calculer les associations SNP-CpG était :

$$\text{mADN} \sim \text{G\_add} + \text{sexe} + \text{gen\_CPs} + \text{comp\_cell} \qquad (3.1)$$

où *G_add* correspond au nombre d'allèles alternatifs (modèle additif : 0, 1 ou 2), *sexe* au sexe de l'enfant, *gen_CPs* aux quatre premières composantes principales de l'analyse en composantes principale faite sur les données génétiques de Gen3G et *comp_cell* à la composition cellulaire estimée sur les données de mADN. Cette dernière se décline en six variables correspondant à des types cellulaires et somment à 1.0 par individu : trophoblastes, stromales, Hofbauer, endotheliales, érythroblastes et syncytiotrophoblastes.

Pour les analyses descriptives, les annotations d'Illumina concernant la position du CpG par rapport aux gènes ou par rapport aux CGI ont été utilisées. Les catégories de position par rapport aux gènes se définissent comme suit : *TSS200* et *TSS1500* pour les CpG situés respectivement à 0–200 et 200-1500 bases en amont du TSS, *5'UTR* pour ceux situés dans la région non traduite en 5' (entre le TSS et le codon de départ ATG), *3'UTR* pour ceux situés dans la région non traduite en 3' (entre le codon stop et la signal poly A), *body* pour ceux situés entre le codon de départ ATG et le codon stop, *1stExon* pour ceux situés dans le premier exon et *ExonBnd* pour ceux situés à une frontière exonique (20 pb en amont ou en aval du début ou de la fin d'un exon). Les catégories de position par rapport aux CGI, quant à elles, sont : *Island* pour les CpG situés dans un CGI, *N_Shore* et *S_Shore* pour ceux situés respectivement de 0à 2 kb en amont (5') ou en aval (3') d'un CGI, que *N_Shelf* et *S_Shelf* pour ceux situés respectivement de 2 à 4 kb en amont (5') ou en aval (3') d'un CGI et *OpenSea* pour ceux à plus de 4 kb d'un CGI.

3.2.6 Croisement avec les tDMR

Le chevauchement entre les mQTL et les tDMR a été fait en utilisant l'outil *intersect* de la suite *bedtools* (Quinlan et Hall, 2010). Les positions des CpG ont été utilisées pour représenter la position des mQTL.

3.2.7 Analyses d'enrichissement de termes ontologiques

Les analyses d'enrichissement de termes ontologiques ont été réalisées avec l'interface web de l'outil *g:Profiler* (Raudvere *et al*., 2019). Les mQTL ayant une valeur *q* inférieure à 1e-75 (N = 2 384) ont été sélectionnés afin de limiter l'analyse aux associations les plus significatives. Cette étape sert également à limiter le nombre de gènes dans l'ensemble pour lequel on souhaite calculer l'enrichissement de termes ontologiques, comme ce type d'analyse fonctionne mieux avec un ensemble de gènes réduit. Cet ensemble de gènes a été construit en utilisant les annotations des CpG fournis par Illumina, comme pour les analyses d'enrichissement de termes ontologiques décrites à la section 2.2.6.4. L'ensemble de gènes de référence a été construit en utilisant ces mêmes annotations, mais pour tous les CpG analysés. Au final, un ensemble de 1803 gènes associés à un CpG d'un mQTL significatif ainsi qu'un ensemble de 25 426 gènes de référence ont été fournis à *g:Profiler*. Les paramètres par défaut ont été utilisés.

3.2.8 Jeu de données de eQTL de placenta utilisé

À partir des mêmes données génétiques que celles utilisées dans ce mémoire et de données d'expression de gènes de la face fœtale du placenta, un autre étudiant du laboratoire, Samuel Côté, a produit des eQTL. Brièvement, l'ARN total a été extrait à l'aide de la trousse d'extraction mirVana Paris Kit de Invitrogen par l'équipe de Gen3G à Chicoutimi à partir des mêmes échantillons de placenta que ceux utilisés pour mesurer la mADN (voir section 3.2.2). Les extractions ont ensuite été envoyées au *Broad Institute*, où la préparation des librairies, le séquençage de l'ARN par RNA-seq et le contrôle de qualité des données ont été effectués. Des données de séquençage ont été générées pour 182 échantillons de la face fœtale du placenta. Le

contrôle de la qualité des lectures, leur alignement et la quantification des transcrits ont été réalisés en utilisant le pipeline de TOPMed RNA-Seq (Broad Institute, n.d.). Un fichier contenant le nombre de lectures par gène et un fichier contenant le niveau d'expression relative de chaque gène normalisé pour la profondeur de séquençage et pour la longueur des transcrits (en TPM) nous ont finalement été fournis. Ces derniers ont été générés par RNA-SeQC v2.3.6 (Graubert *et al*., 2021) avec les annotations de GENCODE version 30 (Frankish *et al*., 2019) pour l'assemblage du génome humain GRCh38. Les lectures ont ensuite été filtrées en suivant le pipeline de GTEx (Aguet *et al*., 2020). Le même pré-traitement sur les données génétiques a été appliqué que celui pour l'Identification des mQTL (voir section 3.2.3). Les eQTL ont été identifiés en utilisant la même fonction de *TensorQTL* que pour l'identification des mQTL, suivant un modèle incluant les six premières composantes principales de la structure populationnelle de Gen3G, 30 variables latentes produites avec l'outil *Peer* (Stegle *et al*., 2012) et l'âge gestationnel.

3.2.9 Croisement avec les résultats de Cardenas *et al.* (2018)

Les valeurs de glycémie maternelle utilisées dans ces analyses sont les mêmes que celles décrites dans Cardenas *et al.* (2018). Brièvement, lors de la visite de suivi au deuxième trimestre de grossesse, les participantes ont passé un test d'hyperglycémie provoquée par voie orale (épreuve d'HGPO), qui consiste à ingérer 75g de sucre en étant à jeun (aussi appelé OGTT, de l'anglais *oral glucose tolerance test*). Leur glycémie a été mesurée à jeun, puis 1h et 2h après l'ingestion du glucose. Les données d'expression utilisées pour ces analyses (gène PDE4B) sont les mêmes que celles ayant été utilisées pour produire les eQTL (voir section 3.2.8). La Figure 3.1 présente la taille des intersections de chacun des ensembles de données. Toutes les analyses d'association ont été faites avec de simples régressions linéaires, sans ajout de covariables, puisqu'il s'agissait seulement d'analyses exploratoires.

Figure 3.1 **Intersection des différents jeux de données utilisés pour l'analyse exploratoire suivant les travaux de Cardenas *et al.* (2018).**

3.3 Résultats et discussion

3.3.1 Description des données

Préalablement au retrait des CpG avec peu de variance, la distribution des données de mADN de placenta demeure cohérente avec celle des tDMR rapportée au Chapitre 2, malgré des changements dans les étapes de prétraitement et l'utilisation d'un sous-ensemble de données différent. Il est cependant possible d'observer une diminution de la proportion des sites fortement méthylés, ce qui pourrait s'expliquer par l'ajout de l'étape de winsorisation des valeurs extrêmes qui a pour conséquence de limiter leur écart avec le groupe et pourrait donc artificiellement diminuer la proportion de sites fortement méthylés. En ne conservant ensuite que les CpG avec une variabilité suffisante pour les analyses d'association (variance supérieure à 1×10^{-5}), on observe principalement une diminution de la proportion des sites avec une faible valeur de mADN (Figure 3.2B). Ceci pourrait s'expliquer par le fait que les CpG avec des valeurs de méthylation extrêmes (y compris les faibles valeurs) ont généralement une plus faible variance (Figure S A.1).

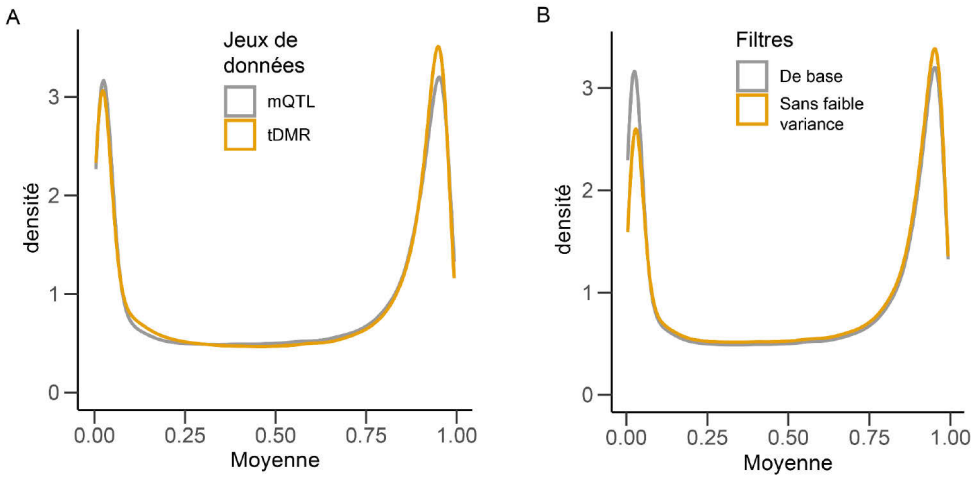

Figure 3.2 **Les niveaux de mADN des jeux de données utilisés pour les analyses de tDMR et de mQTL sont similaires malgré des différences de pré-traitement et le retrait des sondes de faible variance.**

A) Distribution des valeurs de méthylation moyenne dans les jeux de données utilisés pour les analyses de tDMR (jaune) et de mQTL (gris). **B)** Distribution des valeurs de méthylation moyenne dans le jeu de données pour l'analyse de mQTL avant (gris) et après (jaune) avoir filtré pour retirer les sondes de faible variance.

En ce qui concerne les données génétiques, l'analyse de l'ascendance génétique a permis de confirmer que la plupart des participants de Gen3G sont d'ascendance européenne et de retirer dix participants trop éloignés de ce groupe pour pouvoir être inclus dans les analyses statistiques de manière adéquate étant donné leur faible nombre (Figure 3.3). De plus, ces résultats sont cohérents avec l'ethnicité rapportée des participants (données non présentées). L'analyse de parentalité, quant à elle, a permis de retirer 22 individus issus de paires apparentées (coefficient de parenté supérieur ou égal à 2e-5,5, ce qui correspond approximativement à des seconds cousins). Le niveau de parenté parmi ces 22 individus variait de frère-sœur à second cousin. Encore une fois, ces résultats concordent avec les informations filiales auto-rapportées.

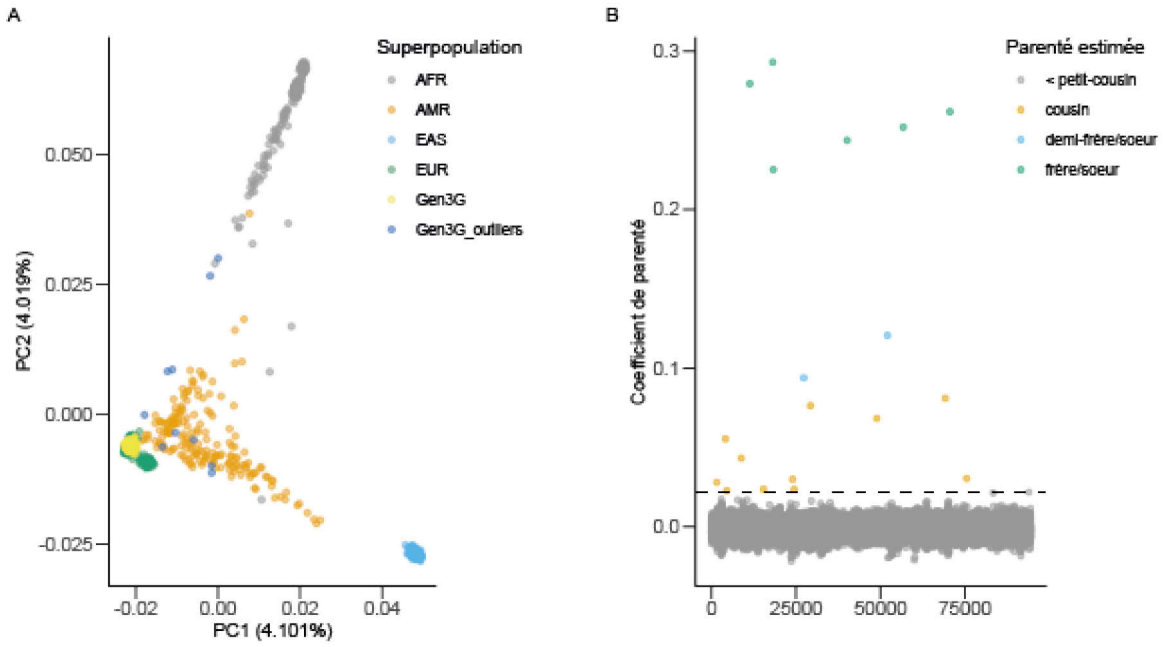

Figure 3.3 **L'ascendance de la majorité des participants de la cohorte Gen3G est européenne et la plupart des individus ne sont pas apparentés, sauf pour 22 d'entre eux.**

A) Estimation de l'ascendance des participants de Gen3G par analyse en composante principale, où chaque point représente un individu. Les codes de superpopulations correspondent à ceux fournis dans la phase 1 du projet *1000 Genomes*. AFR : ascendance africaine, AMR : ascendance américaine, EAS : ascendance est-asiatique, EUR : ascendance européenne. Les cercles bleu foncé identifient les participants de Gen3G retirés en raison d'un trop grand éloignement de l'ascendance européenne. **B)** Estimation de la parentalité des participants de Gen3G, où chaque point représente une paire d'individus. La ligne horizontale représente le seuil de parentalité de 2,2% au-dessus duquel un participant par paire devait être retiré pour cause de trop grande parentalité.

3.3.2 Identification des mQTL

L'objectif principal de ce projet était d'identifier les mQTL dans le placenta afin de les caractériser et de pouvoir les rendre disponibles pour des applications par notre groupe, mais aussi pour d'autres. À partir des données de mADN de la puce EPIC et de données de séquençage de génome entier de 398 participants, nous avons été en mesure d'identifier au moins un mQTL proximal significatif pour 188 529 CpG à travers tout le génome (Figure 3.4A), soit 27,7% des sites testés. Ceci représente une proportion légèrement supérieure à ce qui avait déjà été rapporté (Delahaye *et al.*, 2018). Contrairement à d'autres types d'analyses pangénomiques, par exemple un GWAS cherchant à identifier quelques loci associés à un trait complexe, il est attendu qu'une grande proportion des associations calculées soit significative. Ainsi, on obtient un diagramme quantile-quantile qui serait en d'autres circonstances indicateur de la présence de faux positifs, mais qui est attendu dans des analyses de mQTL (Figure 3.4B).

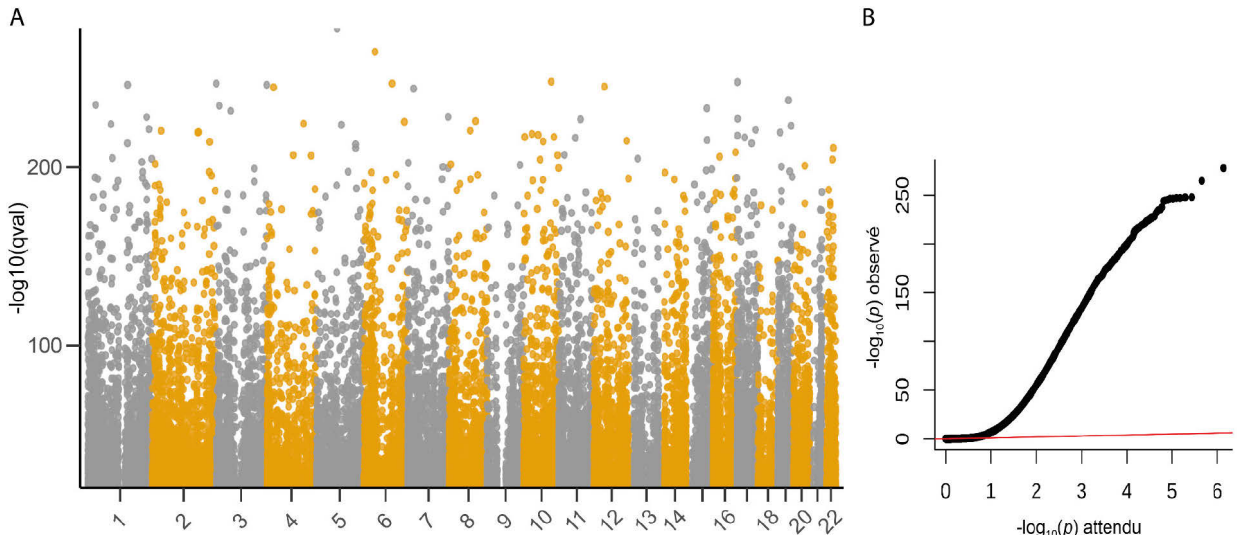

Figure 3.4 **Identification de 188 529 mQTL proximaux dans le placenta.**
A) Diagramme de Manhattan où chaque point illustre la valeur q d'un mQTL significatif (N = 188 529). Les positions sur l'axe des abscisses sont ordonnées et classées par chromosome, mais ne sont pas proportionnelles aux distances réelles. **B)** Diagramme quantile-quantile des valeurs p observées par rapport à celles qui seraient attendues sous l'hypothèse nulle (N = 681 795).

En analysant seulement les mQTL proximaux, on maximise le nombre d'associations trouvées en limitant la perte de puissance due à la correction pour tests multiples. Bien qu'on sacrifie ainsi la découverte d'associations distales, celles-ci ne sont généralement que marginales (par exemple, ~93% des mQTL rapportés par Hannon *et al.* (2016) sont proximaux). Dans nos données, la médiane de la distance absolue entre le SNP et le CpG dans les mQTL était de 14,6 kb et plus de 90% des associations se trouvaient dans une fenêtre de 220 kb, suggérant qu'une fenêtre d'association de 500 kb en amont et en aval était amplement suffisante (Figure 3.5A). De plus, une proportion similaire des SNP se trouvait en amont et en aval des CpG. Il serait raisonnable de réduire cette fenêtre de moitié dans des analyses futures où on voudrait réduire le nombre d'associations à calculer.

Pour chaque CpG, une moyenne de 3 289 SNP proximaux ont été testés, et ces SNP étaient fréquemment testés pour plusieurs CpG. Parmi les 188 529 mQTL, on retrouve 146 878 SNP distincts, indiquant que 22% des CpG partagent le même SNP qu'un autre mQTL. Similairement, parmi les SNP impliqués dans un mQTL, 13% étaient associés à plus d'un CpG (Figure 3.5C). Notons que plusieurs CpG ne partageant pas le même SNP partagent probablement le même locus, c'est-à-dire que les SNP de ces mQTL seraient en déséquilibre de liaison et représenteraient la même association causale cryptique. Il sera intéressant de raffiner ces analyses en groupant entre eux les CpG dont la mADN est corrélée afin de les unifier sous un même phénotype, et d'identifier ainsi les SNP associés à la mADN d'une région. Cette approche, conceptuellement similaire à celle employée pour identifier les tDMR, est proposée dans la suite de *QTLtools* (Delaneau *et al*., 2017).

En ce qui concerne le sens des associations, la présence de l'allèle alternatif est légèrement plus fréquemment associée à la baisse de la mADN qu'à son augmentation (respectivement 53% et 47%, Figure 3.5B). Il semble de plus exister dans ce groupe un sous-groupe d'associations se distinguant dans leur association entre la pente et la significativité de la relation, formant un « pétale » à gauche du diagramme en volcan. Après investigation, ces associations sont principalement celles où le SNP se trouve à la position du CpG (0,16% de tous les mQTL et

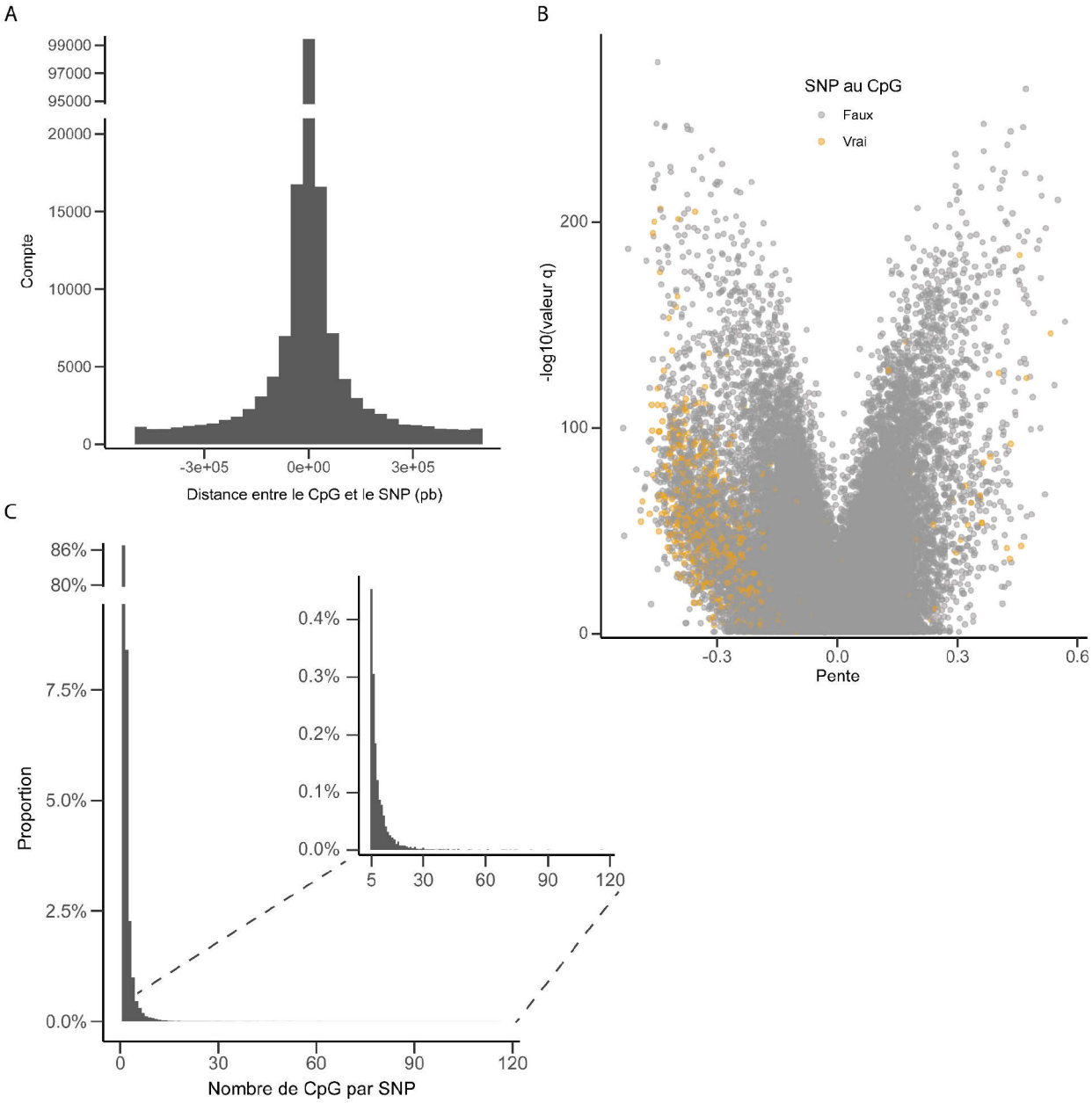

Figure 3.5 **Position des SNP par rapport aux mQTL.**

A) Distribution de la distance entre les CpG et les SNP pour tous les mQTL significatifs (N = 188 529). **B)** Diagramme en volcan où chaque point représente un mQTL (N = 681 795). Les points en jaune sont les mQTL pour lesquels le SNP est à la position du site CpG (N = 1 108). **C)** Distribution du nombre d'association mQTL significative par SNP.

0,56% des mQTL significatifs); il est alors évident que la présence de l'allèle alternatif soit associée à une baisse de la mADN mesurée à ce site. Dans les étapes de pré-traitement des données de mADN, les CpG ayant un SNP avec une fréquence allélique supérieure à 5% (selon le fichier d'annotation d'Illumina) dans leur séquence cible ou au nucléotide d'extension ont été retirés, laissant dans l'ensemble de données ceux pour lesquels la fréquence allélique du SNP est inférieure 5%, ou alors plus grande dans notre cohorte que dans la population de référence utilisée par Illumina. Si ces mQTL devaient être utilisés dans des analyses subséquentes, ils devraient être traités avec une attention particulière puisque l'effet de la présence de l'allèle alternatif et celui de la mADN pourraient être confondus. Dans le futur, il pourrait être intéressant de refaire certaines analyses présentées plus loin en retirant ces mQTL afin d'en estimer l'impact.

Si on s'intéresse maintenant au contexte génomique des mQTL, il est possible de constater que moins de CpG impliqués que non impliqués dans un mQTL se trouvent dans des îlots (Figure 3.6A, catégories "Island" et "OpenSea"), ce que d'autres avaient déjà rapporté (Oliva *et al*., 2022). Parmi les CpG étant dans un mQTL significatif, on observe aussi une association moins fréquente à un gène (Figure 3.6B, catégorie "None"). Ceci pourrait s'expliquer par une régulation plus fine de la mADN à proximité des gènes, c'est-à-dire que davantage d'éléments contribueraient à la régulation de la mADN et auraient donc individuellement un moins grand effet. On détecterait alors moins de mQTL dans ces régions parce que la taille d'effet du SNP serait généralement plus petite et nécessiterait plus de puissance statistique pour les détecter. Cette hypothèse n'a pas été testée dans le cadre de ce travail, mais présente une piste intéressante pour la suite des travaux.

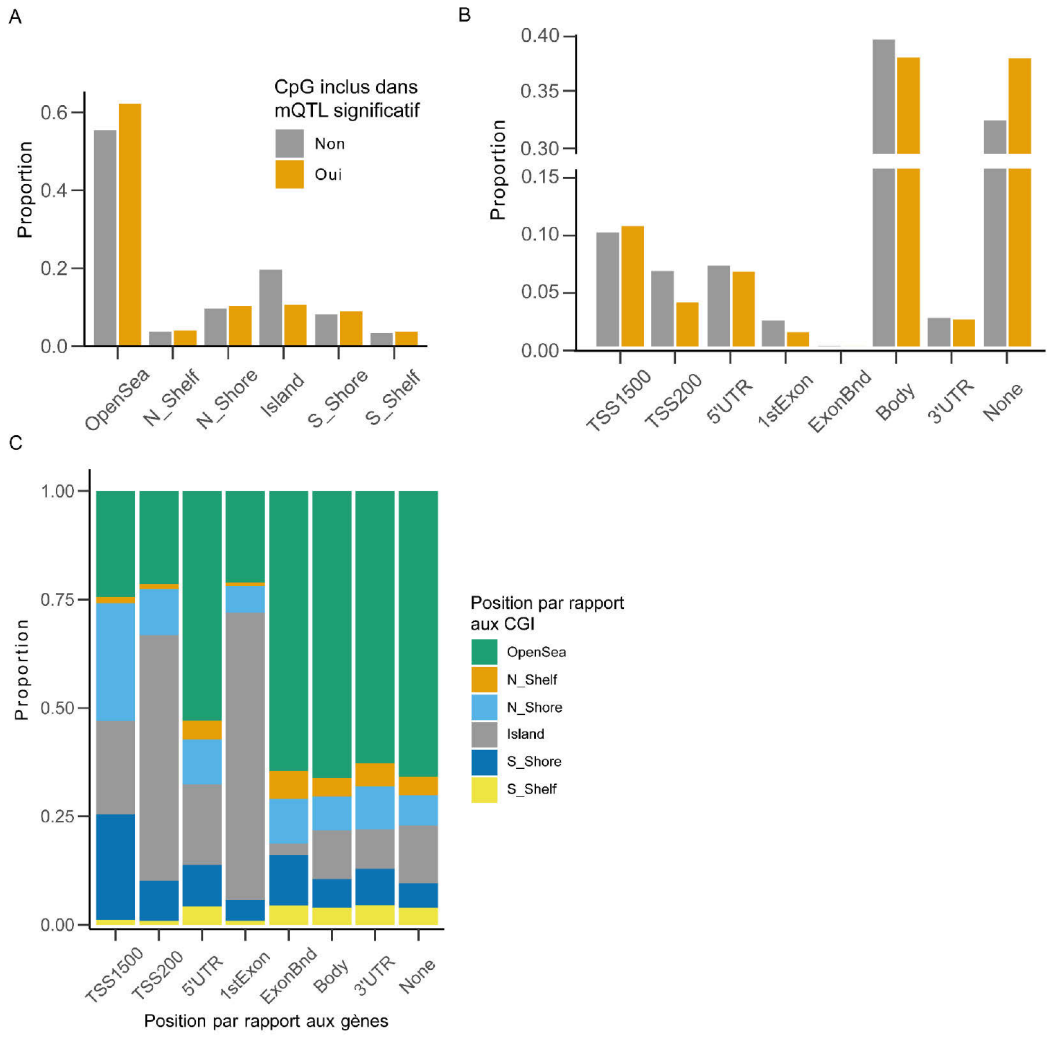

Figure 3.6 **Les CpG des mQTL significatifs sont plus éloignés des CGI et davantage situés des régions intergéniques que les CpG non associés à des mQTL.** Les barres jaunes représentent les CpG associés à des mQTL significatifs, alors que les barres grises représentent la fraction complémentaire qui ne l'est pas. **A)** La proportion des CpG associés à chacune des positions relatives aux CGI, selon les annotations d'Illumina. **B)** La proportion des CpG associés à chacune des positions relatives aux gènes, selon les annotations d'Illumina. Quand plus d'une annotation était donnée pour un CpG, la plus fréquente a été retenue. **C)** Croisement entre les positions relatives aux CGI et aux gènes de tous les CpG inclus dans l'analyse.

Il semble également y avoir une relation entre la pente de l'association mQTL et la position du CpG par rapport aux CGI ou par rapport aux gènes. On observe en effet que la distribution des pentes est bimodale pour la plupart des catégories de position des CpG avec un mode de part et d'autre du zéro (Figure 3.7A-B). Les catégories faisant exception sont celles des CGI, du premier exon, des frontières exoniques et du TSS200, ayant tous un seul mode plus grand que 0, donc généralement une corrélation positive entre la mADN et la présence de l'allèle alternatif. Ces résultats sont cohérents entre eux puisqu'on observe que les catégories du premier exon et du TSS200 sont particulièrement riches en CGI (Figure 3.6C). La distribution des pentes observée pour la catégorie des frontières exoniques pourrait représenter un artéfact dû au faible nombre de CpG qu'elle représente par rapport à celles du premier exon ou du TSS200 (155, 13 293 et 39 437 CpG, respectivement). La pente positive des mQTL dans ces régions pourrait s'expliquer par une baisse de l'affinité des facteurs de transcription à l'ADN. En effet, si ces SNP situés autour d'un promoteur perturbent des motifs de liaison à l'ADN, le recrutement de la machinerie transcriptionnelle pourrait être affecté. Comme on sait qu'au niveau des promoteurs transcriptionnellement actifs la chromatine est remodelée pour présenter, entre autres, moins de mADN, ce recrutement moins efficace pourrait se traduire par des taux plus élevés de mADN. La première étape pour tester cette hypothèse pourrait être d'analyser les séquences dans lesquelles se trouvent les SNP afin de vérifier si elles sont enrichies en motifs de liaison à l'ADN reconnus par des facteurs de transcription.

3.3.3 Croisement avec d'autres ensembles de mQTL

Dans leurs travaux s'intéressant à la relation entre les variants génétiques, la mADN dans le placenta ainsi qu'à l'expression des gènes dans ce même tissu, Delahaye *et al.*(2018) rapportent 4 342 mQTL. En croisant ces derniers avec les mQTL présentés dans ce mémoire, on retrouve un chevauchement de 2 968 CpGs, mais seulement 361 sont associés au même SNP principal. Parmi ceux n'étant pas associés au même SNP, il était intéressant de vérifier si les SNP étaient rapprochés. Cette proximité a été établie à moins de 5 kb, ce qui augmenterait les chances que les SNP fassent partie d'un même locus (Shifman *et al*., 2003). Un total de 1 136 SNP (43,6% des non concordants) se trouvaient à une distance de moins de 5 kb. En conséquence, environ

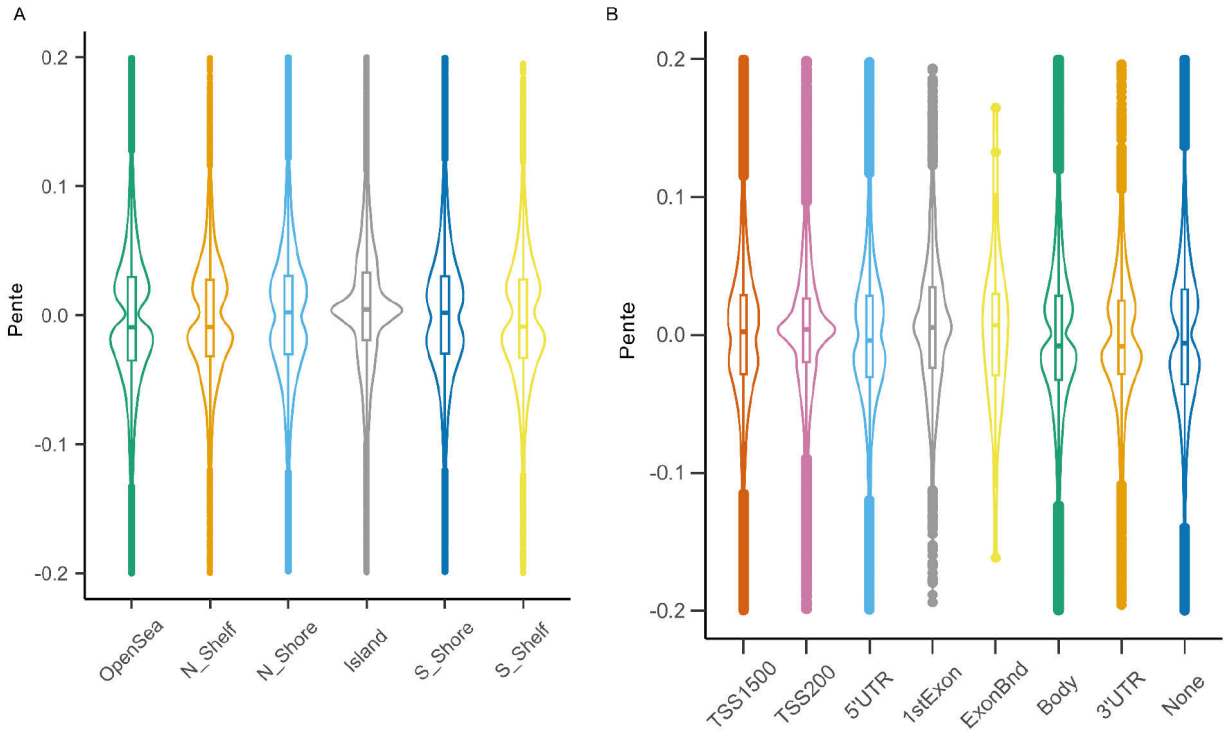

Figure 3.7 **Les pentes des associations mQTL significatives varient en fonction de la position par rapport aux gènes ou par rapport aux CGI.**

Les axes des ordonnées ont été limités à -0,2 et 0,2 afin de mieux visualiser les tendances centrales. **A)** Distributions des pentes des mQTL significatifs en fonction de la position par rapport aux CGI, selon les annotations d'Illumina. **B)** Distributions des pentes des mQTL significatifs en fonction de la position par rapport aux gènes, selon les annotations d'Illumina. Quand plus d'une annotation était donnée un CpG, la plus fréquente a été retenue.

la moitié des CpG communs étaient associés au même locus dans nos résultats et dans ceux de Delahaye *et al.* (2018). De plus, la plupart des mQTL (81,3%) avaient la même direction de pente, indépendamment de s'ils impliquaient le même SNP ou non. Considérant les différences statistiques et méthodologiques, le recoupement observé entre ces deux ensembles de mQTL augmente la crédibilité de nos résultats. Delahaye *et al.* (2018) rapportent aussi le croisement de leur propre mQTL avec ceux de Do *et al.*, pour un total de 327 même mQTL. Parmi ceux-ci,

on retrouve un chevauchement de 255 CpG avec ceux des analyses présentées dans ce mémoire, dont 33 ont le même SNP significatif. Parmi ceux n'étant pas associés au même SNP, 118 (53,2% des non concordants) se trouvent à une distance d'au plus 5 kb, pour une concordance globale de locus d'environ 60%.

3.3.4 Croisement avec les tDMR

Un total de 7 320 CpG de mQTL significatifs chevauchent avec un total de 2 837 des régions tDMR ayant passé les filtres décrits au Chapitre 2. Comme on suppose que les valeurs de mADN des CpG au sein d'un tDMR sont corrélées, on pouvait poser l'hypothèse que les mQTL d'une même tDMR étaient associés à un SNP commun. Il y avait effectivement 323 groupes de CpG (plus de 3 CpG) qui étaient dans un même tDMR et partageaient le même SNP. Il ne semblait cependant pas y avoir de relation entre le sens de l'association mQTL et le sens des tDMR, ni d'enrichissement pour un sens de pente de mQTL.

3.3.5 Enrichissement de termes ontologiques

Une analyse d'enrichissement des termes ontologiques des gènes associés aux CpG des mQTL significatifs a été faite avec l'outil en ligne *g:profiler* (pour les résultats complets, consulter https://biit.cs.ut.ee/gplink/l/xa4Nbu7oR-). En plus des résultats classiques d'enrichissement de termes ontologiques, cet outil met également en évidence les termes principaux (*driver terms*) dans chaque catégorie de termes ontologiques du *Gene Ontology Consortium*. En résumé, ces derniers ont pour objectif de représenter les termes enrichis sur la base des gènes ayant contribué à l'enrichissement des termes regroupés en sous-ontologies (Alexa *et al*., 2006; g:Profiler, n.d.).

Sur la base de ces termes principaux, on remarque pour les fonctions moléculaires un enrichissement se rapportant à la liaison du phosphatidylinositol, à l'activité des pompes à ion calcium et à la liaison des ions métalliques (Figure 3.8, Tableau 3.2). Pour les processus moléculaires, neuf termes principaux sont identifiés, soit les processus métaboliques du phosphore, des composés organo-azotés, des petites molécules et dérivé des glucides, en plus

des processus cataboliques, du système circulatoire et de biosynthèse des glucides ainsi que le transport des ions métalliques et la régulation de la thermogenèse induite par le froid (Figure 3.8, Tableau 3.2). De plus, bien que n'étant pas des termes principaux, plusieurs termes se rapportant au développement sont enrichis, par exemple, le développement d'organisme multicellulaire et la morphogenèse des structures anatomiques.

Les termes basés sur le *Kyoto Encyclopedia of Genes and Genomes* (KEGG) *PATHWAY* soulignent quant à eux, entre autres, les voies de signalisation de l'ocytocine, de la relaxine et de l'œstrogène, la synthèse, sécrétion et action d'hormones de croissance, la synthèse et la sécrétion de l'aldostérone ainsi que la résistance à l'insuline (Tableau 3.2).

Il est intéressant de constater que plusieurs des termes enrichis concordent avec les fonctions placentaires connues. En effet, on peut supposer que les fonctions de filtrations et d'échange de nutriments du placenta sont mises en évidence par les termes en lien avec la liaison et le transport de petites molécules (par exemple, GO:0015085, GO:0046872, GO:0044281 et GO:0030001). Dans le même ordre d'idées, l'importance du placenta dans la régulation de l'apport en glucides du fœtus semble se refléter dans les termes GO:1901135 et GO:0016051, d'autant plus que l'on sait que le placenta synthétise certains glucides simples (Maltepe et Fisher, 2015). Les termes KEGG semblent quant à eux mettre la lumière sur les fonctions endocriniennes et métaboliques du placenta, notamment par la production d'œstrogène et d'hormones de croissance (Costa, 2016; Gimpl et Fahrenholz, 2001; Maltepe et Fisher, 2015), mais aussi par le rôle qu'il joue dans le développement d'une certaine résistance à l'insuline normale durant la grossesse (Costa, 2016).

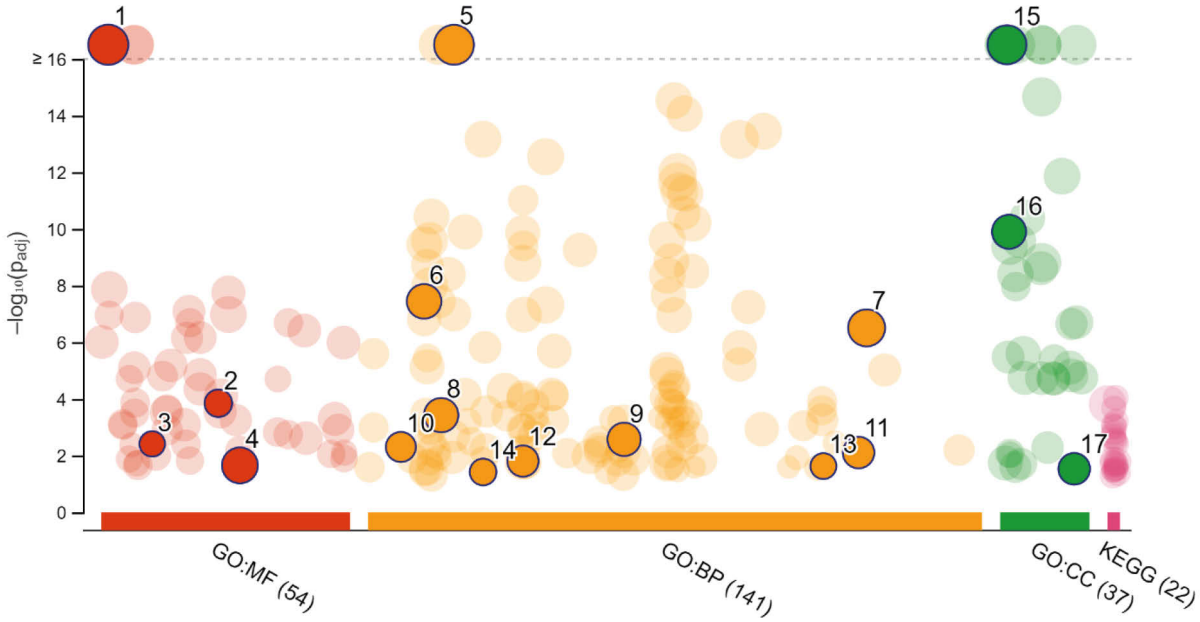

Figure 3.8 **Enrichissement des termes ontologiques associés aux mQTL significatifs.** Dans le graphique, les valeurs -log10(p) ont été bornées à 16 afin de mieux visualiser les plus petites valeurs. Les termes ontologiques numérotés représentent les termes principaux et peuvent être retrouvés au Tableau 3.X. L'image a été modifiée de la page de survol des résultats générée par g:profiler.

Tableau 3.2 Enrichissement des termes ontologiques associés aux mQTL significatifs.

ID	Source	ID du terme	Terme	Valeur *p* ajustée
1	GO:MF	GO:0003674	molecular_function	9,15 x 10-47
2		GO:0035091	phosphatidylinositol binding	1,43 x 10-04
3		GO:0015085	calcium ion transmembrane transporter activity	3,88 x 10-03
4		GO:0046872	metal ion binding	2,23 x 10-02
5	GO:BP	GO:0009987	cellular process	1,05 x 10-27
6		GO:0006793	phosphorus metabolic process	3,61 x 10-08
7		GO:1901564	organonitrogen compound metabolic process	3,11 x 10-07
8		GO:0009056	catabolic process	3,75 x 10-04
9		GO:0044281	small molecule metabolic process	2,68 x 10-03
10		GO:0003013	circulatory system process	4,98 x 10-03
11		GO:1901135	carbohydrate derivative metabolic process	7,84 x 10-03
12		GO:0030001	metal ion transport	1,57 x 10-02
13		GO:0120161	regulation of cold-induced thermogenesis	2,36 x 10-02
14		GO:0016051	carbohydrate biosynthetic process	3,80 x 10-02
15	GO:CC	GO:0005737	cytoplasm	5,46 x 10-34
16		GO:0005856	cytoskeleton	1,26 x 10-10
17		GO:0099081	supramolecular polymer	2,94 x 10-02
NA	KEGG	KEGG:04921	Oxytocin signaling pathway	9,21 x 10-05
NA		KEGG:04926	Relaxin signaling pathway	1,97 x 10-02
NA		KEGG:04915	Estrogen signaling pathway	2,92 x 10-02
NA		KEGG:04935	Growth hormone synthesis, secretion and action	6,18 x 10-04
NA		KEGG:04925	Aldosterone synthesis and secretion	2,94 x 10-02
NA		KEGG:04931	Insulin resistance	2,21 x 10-02

3.3.6 Exemple d'utilisation des mQTL hors-Gen3G

Comme mentionné au Chapitre 1, les mQTL produits par une étude peuvent parfois être utilisés par un autre groupe afin de nourrir des hypothèses au sujet des mécanismes biologiques derrière les résultats de l'étude. Les mQTL présentés dans ce mémoire ont pu être utilisés à cet effet dans une étude visant à identifier les déterminants génétiques du poids du placenta, ainsi qu'à départager l'influence du génome maternel, paternel et fœtal sur cette variable. L'article a été accepté par le journal Nature Genetics et sera éventuellement publié, mais peut être présentement trouvé sous sa forme non révisée sur MedRxiv (Beaumont *et al*., 2022).

En bref, une méta-analyse de 28 cohortes de participants d'ascendances européennes a été réalisée, permettant l'identification de 40 loci associés au poids du placenta. Les auteurs ont croisé les mQTL présentés dans ce mémoire avec leurs résultats de GWAS sur la base de SNP identiques, associant ainsi 21 des 40 loci à un mQTL. Pour trois de ces mQTL, le CpG est situé dans des sites d'hypersensibilité à la DNAse 1 spécifiques à des tissus relatifs au développement. Comme il s'agissait d'une étude à couverture large, les auteurs n'ont pas davantage élaboré la piste des mQTL. On peut cependant supposer que la concordance de leur SNP d'intérêt avec des mQTL situés dans des sites d'hypersensibilité à la DNAse 1, donc associées à des régions transcriptionnellement actives, donne davantage de crédibilité aux associations trouvées et offre des pistes à explorer afin d'expliquer les mécanismes sous-jacents.

3.3.7 Exemple d'utilisation des mQTL intra-Gen3G

Parmi les études réalisées avec les données de Gen3G, certaines se sont penchées sur l'association entre le mADN et des traits relatifs à la grossesse. Comme expliqué au Chapitre 1 et exemplifié par l'étude de Beaumont *et al.*, il est possible de tirer profit des mQTL pour pousser plus loin ce type d'analyse d'association. C'est donc ce que j'ai voulu faire en revisitant les résultats d'un article publié en 2018 par Cardenas *et al.* (2018) sur la réponse du méthylome placentaire à la glycémie en grossesse.

3.3.7.1 Mise en contexte

Brièvement, le but de l'étude était d'explorer le rôle du placenta dans la relation déjà connue entre la dérégulation de la glycémie maternelle et les effets délétères sur la santé de l'enfant (par exemple, l'hypoglycémie néonatale). Les auteurs ont identifié sept sites CpG qui seraient associés à la glycémie maternelle 2h après l'épreuve d'HGPO (simplement « HGPO » pour le reste du texte), dont quatre se trouvent dans un locus du gène de *PDE4B* (Tableau 3.3, Figure 3.9). Ils rapportent également une corrélation positive significative entre le niveau d'expression de ce gène et la mADN pour un des quatre CpG du locus (cg03442467). Leur hypothèse est que l'augmentation de la mADN au CpG cg03442467 serait un mécanisme d'adaptation placentaire à l'hyperglycémie maternelle. On suppose que l'augmentation de l'expression de *PDE4B* serait une composante de l'adaptation à l'hyperglycémie, mais les auteurs ne font pas mention de cette analyse ou de cette extension de l'hypothèse.

Les auteurs mentionnent toutefois qu'avec les données à leur disposition, il n'était pas possible de tester le sens de l'association entre l'HGPO et la mADN. En plus de l'hypothèse mise de l'avant dans l'article, il existe deux autres possibilités : un troisième facteur pourrait à la fois causer l'augmentation de l'hyperglycémie maternelle et la mADN au locus de *PDE4B*, ou encore l'augmentation de la mADN pourrait faire partie de la cascade d'événements menant à la dérégulation de la glycémie. Mon hypothèse était que les mQTL pourraient nous informer sur le scénario le plus probable si on utilisait des approches statistiques de la famille de l'inférence causale.

Notons que les données de mADN utilisées dans l'article de Cardenas *et al.* (2018) sont les mêmes que celles utilisées au Chapitre 2, incluant les étapes de pré-traitement. Les mesures d'HGPO utilisées sont également les mêmes. Seules les données d'expression diffèrent car, pour l'article, elles ont été produites par PCR quantitatif en temps réel alors que celles utilisées pour les analyses présentées plus bas proviennent de séquençage d'ARN (voir les sections 3.2.8 et 3.2.9).

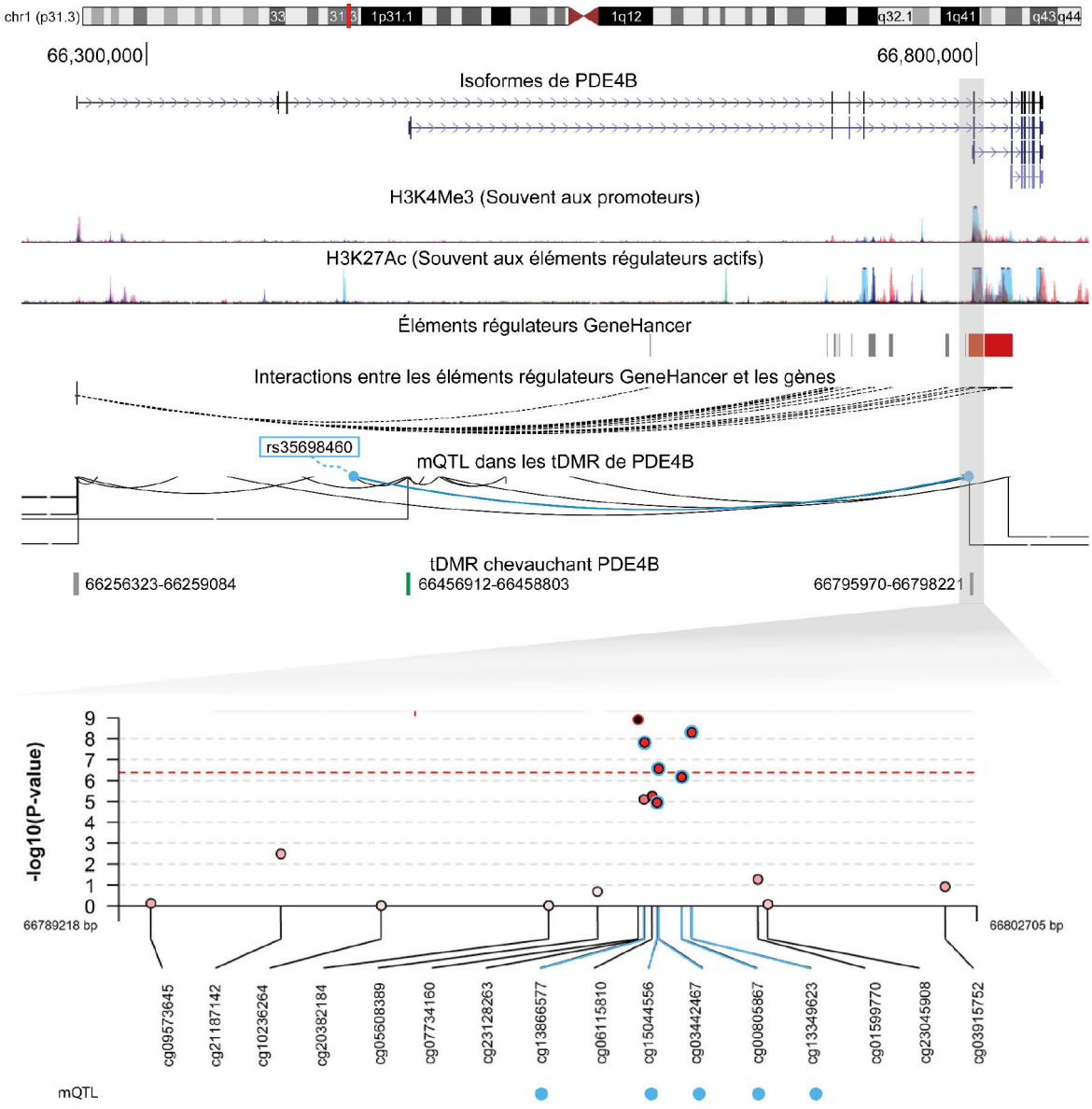

Figure 3.9 **Le locus présenté dans l'article de Cardenas *et al.* (2018) contient plusieurs CpG associés au SNP rs35698460.**

La figure a été générée avec le *Genome Browser* (partie du haut) et en modifiant la Figure 2 de Cardenas *et al.* (2018) (partie du bas) présentant la relation entre des CpG et l'HGPO. Les pistes présentant les mQTL et les tDMR ont été produites à partir des données présentées dans ce mémoire. La couleur des tDMR représente si la région a passé (vert) ou non (gris) les filtres de significativité présentés au Chapitre 2. Les mQTL mis en évidence en bleu sont ceux du cluster de Cardenas *et al.* (2018) impliquant le SNP rs35698460.

3.3.7.2 Croisement avec les mQTL

On observe dans le Tableau 3.3 ainsi que dans la Figure 3.9 que plusieurs CpG du locus rapporté dans l'article sont associés au même SNP, soit le rs35698460 où une thymine est remplacée par une cytosine. La fréquence de c*et al*lèle dans notre cohorte (24,5%) est similaire à celle attendue dans une population d'ascendance européenne (24,85%), selon *dbSNP* (https://www.ncbi.nlm.nih.gov/snp/rs35698460). Ce SNP n'a jamais été rapporté dans *ClinVar* ni encore comme QTL dans le portail de GTEx. Pour tous les CpG testés (les cinq CpG de l'article de Cardenas *et al.* (2018) qui étaient aussi inclus dans l'analyse de mQTL), la présence de l'allèle alternatif était associée à une augmentation de la mADN (Tableau 3.3, Figure 3.10).

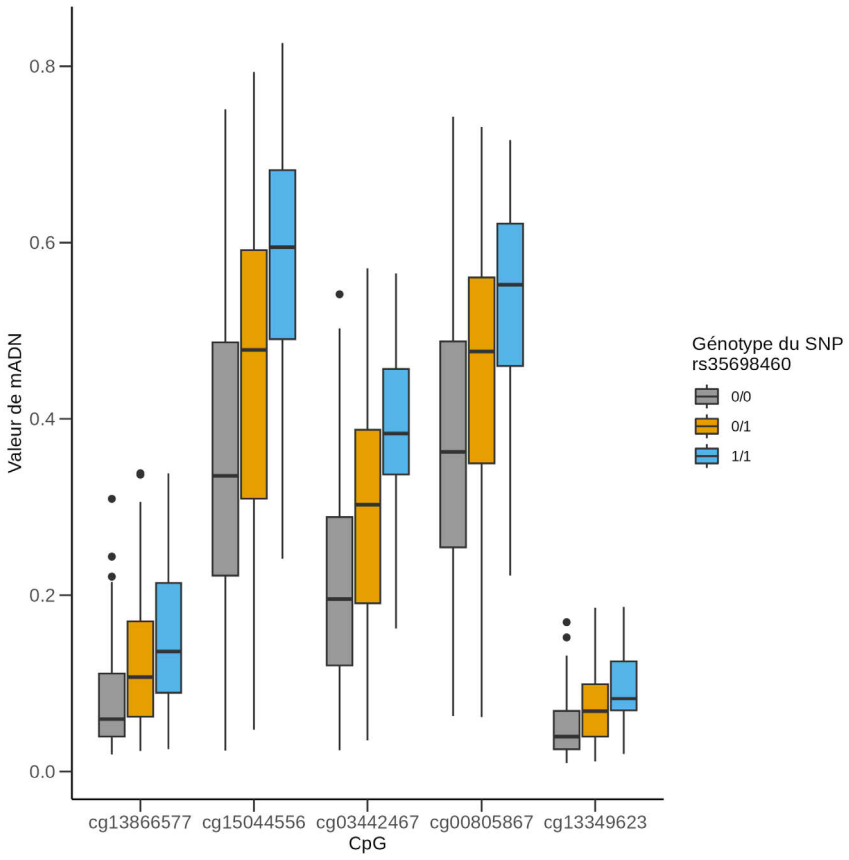

Figure 3.10 La mADN aux CpG d'intérêt identifiés par Cardenas *et al.* (2018) est corrélée avec le génotype du SNP rs35698460.

Tableau 3.3 **Résumés des travaux de Cardenas *et al.* (2018) pour le gène *PDE4B* et leur croisement avec les mQTL.**

			Cardenas *et al*		Ce mémoire					
					mQTL (rs35698460)		HGPO		Expression de PDE4B	
CpG	Position (chr1)	Inclus dans projet mQTL	HGPO $-\log_{10}(p)$	Expres. PDE4B $-\log_{10}(p)$	$-\log_{10}(q)$	pente	$-\log_{10}(p)$	pente	$-\log_{10}(p)$	pente
cg07734160	66797378	Non	8,92	1,05	-	-	-	-	-	-
cg23128263	66797473	Non	-	-	-	-	-	-	-	-
cg13866577	66797481	Oui	6,95	0,57	6,89	0,039	3,74	$-8{,}22 \times 10^{-3}$	0,74	1,609
cg06115810	66797599	Non	-	-	-	-	-	-	-	-
cg15044556	66797677	Oui	-	-	8,47	0,1054	3,87	$-2{,}38 \times 10^{-2}$	0,89	0,658
cg03442467	66797701	Oui	6,55	2,74	11,59	0,0841	4,56	$-1{,}80 \times 10^{-2}$	0,93	0,990
cg00805867	66798063	Oui	-	-	6,09	0,0813	3,93	$-2{,}06 \times 10^{-2}$	0,68	0,646
cg13349623	66798221	Oui	8,69	1,05	7,7	0,0224	4,28	$-5{,}19 \times 10^{-3}$	0,7	2,765

3.3.7.3 Croisement avec les tDMR

Puisque plusieurs CpG consécutifs étaient associés au même SNP, il paraissait intéressant de vérifier si ceux-ci se trouvaient dans un tDMR. En ne prenant que l'ensemble de tDMR filtrés (voir section 2.2.5), il n'existait qu'une seule région associée à *PDE4B* et cette dernière ne correspondait pas au locus identifié dans l'article de Cardenas *et al.* (2018). Ce tDMR comportait neuf CpG, avait une différence moyenne de mADN avec le sang de cordon de -0,444 et s'étendait sur 1 891 pb (chr1:66456912-66458803) dans le corps du gène de *PDE4B* (Figure 3.9, Tableau 3.4). Six des neuf CpG de cette région étaient retrouvés dans une association de mQTL significative et trois de ces six mQTL impliquaient le SNP rs35698460 (Tableau 3.4). En retirant le filtre sur la différence de valeur de mADN, on retrouvait alors deux régions supplémentaires associées à *PDE4B* (Figure 3.9, Tableau 3.4), dont un tDMR de 2 251 pb (chr1:66795970-66798221) englobant sept CpG et chevauchant la région d'intérêt identifiée dans l'article de Cardenas *et al.* (2018). Cinq de ces sept CpG étaient significativement associés au SNP rs35698460.

3.3.7.4 Croisement avec les eQTL et les données d'expression

En cherchant parmi les eQTLs préliminaires de la face fœtale du placenta, aucune association entre l'expression de *PDE4B* et un SNP n'a été trouvée. Notamment, il n'y avait aucune corrélation entre le génotype du SNP rs35698460 et les niveaux d'expression de *PDE4B*. Il n'a pas non plus été possible de répliquer l'association entre l'expression de *PDE4B* et le CpG cg03442467 (R^2 = 0,011, p-value > 0,116), ni aucun autre CpG du locus de la Figure 3.9 (Tableau 3.3). Il n'est cependant pas si surprenant que ces résultats ne soient pas répliqués étant donné de la différence méthodologique importante pour la mesure du niveau d'expression de *PDE4B*. De plus, ces analyses d'associations entre la mADN et l'expression de *PDE4B* ont été faites sans ajouter aucune covariable. Elles pourraient être reproduites en utilisant les mêmes modèles statistiques que ceux de Cardenas *et al.* (2018), soit en incluant des covariables et en utilisant des valeurs M pour représenter la mADN.

Tableau 3.4 Intersection des mQTL et des tDMR associés au gène *PDE4B*

tDMR				CpG		mQTL		
Début (chr1)	**Fin (chr1)**	**Nb CpG**	**Δ mADN**	**ID**	**Pos (chr1)**	**Pente**	**Signif**	**SNP**
66256323	66259084	11	0,047	cg07914107	66256911	-0,038	Non	rs12038590
				cg22733910	66257553	-0,003	Non	rs6698162
				cg04704294	66257822	0,035	Oui	rs4655801
				cg16294013	66258022	0,077	Non	rs775907428
				cg22336004	66258035	0,056	Non	rs775907428
				cg22488256	66258046	0,069	Non	rs1200561783
				cg10552807	66258049	0,114	Non	rs775907428
				cg09444657	66258196	0,138	Non	rs775907428
				cg26832142	66258441	0,057	Non	rs201551177
				cg00046625	66258760	0,047	Non	rs17127920
				cg26963271	66259081	-0,027	Non	rs61781378
				cg24637364	66259084	0,086	Non	rs146376555
66456912	66458803	9	-0,444	cg17590234	66456912	0,026	Oui	rs35185259
				cg20612286	66457146	0,018	Non	rs35297113
				cg11741987	66457416	-0,010	Oui	rs11208724
				cg17563620	**66457970**	**-0,049**	**Oui**	**rs35698460**
				cg08462879	**66458116**	**-0,034**	**Oui**	**rs35698460**
				cg03435439	66458361	-0,062	Oui	rs61799586
				cg26912671	**66458803**	**-0,082**	**Oui**	**rs35698460**
66795970	66798221	7	0,247	cg20382184	66795970	-0,007	Non	rs2312589
				cg05608389	66796739	0,016	Non	rs141859913
				cg13866577	**66797481**	**0,039**	**Oui**	**rs35698460**
				cg15044556	**66797677**	**0,105**	**Oui**	**rs35698460**
				cg03442467	**66797701**	**0,084**	**Oui**	**rs35698460**
				cg00805867	**66798063**	**0,081**	**Oui**	**rs35698460**
				cg13349623	**66798221**	**0,022**	**Oui**	**rs35698460**

En ce qui concerne l'apparente absence de corrélation entre le génotype de rs35698460 et l'expression de *PDE4B*, elle n'invalide pas la possibilité qu'il existe une relation (non détectée dans cette analyse) entre la mADN au CpG cg03442467 et l'expression du gène, même si ce SNP et ce CpG sont dans une relation mQTL. En effet, une multitude d'autres facteurs que le génotype influencent la mADN ou l'expression d'un gène, comme la présence d'autres marques épigénétiques. Dans le cas présent, la présence de l'allèle alternatif semble augmenter la mADN (Figure 3.10), mais le mécanisme sous-jacent pourrait être relativement indépendant des autres processus associés à l'activité transcriptionnelle.

3.3.7.5 Croisement avec les résultats de l'HGPO

Comme on pouvait s'y attendre étant donné l'utilisation des mêmes données, il a été possible de répliquer toutes les associations entre la mADN et les résultats d'HGPO pour les CpG qui se trouvaient dans le jeu de données des mQTL, c'est-à-dire les CpG cg13866577, cg15044556, cg03442467, cg00805867 et cg13349623 (Tableau 3.2), et ce, même si le modèle de régression utilisé n'incluait aucune covariable et utilisait des valeurs bêta pour représenter la mADN. Le génotype du SNP rs35698460, auquel sont associés tous les CpG listés ci-haut, n'est cependant pas corrélé avec l'HGPO, pas plus que les niveaux d'expression de *PDE4B*. Encore une fois cependant, ces analyses ont été réalisées en n'incluant aucune covariable et devraient être refaite de manière plus robuste avant de s'avancer dans davantage d'interprétations.

En somme, cette analyse exploratoire où l'on tentait d'utiliser un mQTL pour déduire le sens de la corrélation entre la mADN et deux traits (la glycémie ainsi que les résultats d'HGPO) n'a pas été aussi informative qu'espéré. Le SNP rs35698460 ne s'est pas avéré pouvoir être utilisé en tant que variable instrumentale dans la corrélation entre le cg03442467 et l'expression de *PDE4B* ou le résultat de l'HGPO, donc il n'a pas été possible d'aller de l'avant avec des méthodes d'inférence causale. Il est raisonnable de croire que, s'il existe une telle variable, sa taille d'effet soit trop petite pour être détectée avec la taille de nos jeux de données. La question demeure donc en suspend jusqu'à ce qu'elle puisse être réabordée avec davantage d'échantillons et des modèles statistiques plus appropriés, notamment par l'inclusion de covariables.

CHAPITRE 4
DISCUSSION ET CONCLUSION

4.1 Discussion générale

Le projet de recherche présenté dans ce mémoire avait comme visée première de contribuer à la caractérisation et la compréhension du méthylome placentaire. Pour ce faire, deux objectifs ont été proposés et atteints : comparer les méthylomes placentaires et de sang de cordon ombilical ainsi que cartographier et caractériser les mQTL de placenta.

Concernant le premier objectif, dont les résultats ont fait l'objet d'une publication, nous avons montré que la majorité des régions différemment méthylées entre le placenta et le sang de cordon ombilical (tDMR) sont moins méthylées dans le placenta, et que les termes ontologiques en lien avec ces régions sont enrichies pour des fonctions spécifiques à chacun de ces tissus, comme celles du système immunitaire pour le sang de cordon et celles liées à la régulation des micro-ARN pour le placenta, ce qui est cohérent avec les récents résultats de (Rondinone *et al*., 2021). De plus, les tDMR moins méthylés dans le placenta étaient associés à des gènes et pseudogènes de la famille des récepteurs olfactifs. En somme, les tDMR identifiés soulignent des régions génomiques liées aux rôles distincts du placenta et du sang de cordon et, en ce sens, contribuent à mieux comprendre le méthylome placentaire par rapport à celui d'un autre tissu fœtal.

Pour ce qui est du deuxième objectif, il a été possible d'identifier 188 529 mQTL proximaux à travers tout le génome, principalement non associés à des gènes ou encore dans le corps de ceux-ci. Les termes ontologiques étaient enrichis pour des fonctions connues du placenta, comme le transport de petites molécules, les processus biologiques des glucides ainsi que la synthèse et la sécrétion d'œstrogène et d'hormones de croissance. Les mQTL ont aussi pu être utilisés dans un projet en cours de publication hors Gen3G ainsi que mis à profit pour explorer sous un nouvel angle les résultats publiés de Cardenas *et al.* (2018). Dans ce dernier volet, il a été observé que plusieurs CpG que les auteurs rapportaient associés à la glycémie maternelle provoquée étaient

impliqués dans des mQTL associés à un SNP, le rs35698460, bien que ce dernier ne soit pas lui-même associé à la glycémie. Ainsi, presque tous les objectifs initiaux ont pu être atteints, sauf le partage des mQTL. Cette dernière étape pourra être franchie au moment de produire une publication rapportant, entre autres, les résultats présentés au Chapitre 3.

Un aspect récurrent dans la littérature de la mADN placentaire n'a pas été abordé dans le présent travail : celui des transposons. On sait par exemple que certains gènes d'enveloppe virale incorporés dans le génome humain sont nécessaires à la placentation, comme les syncytines, et que leurs niveaux de mADN sont plus faibles dans le placenta que dans les autres tissus embryonnaires afin de permettre leur expression (Bolze *et al*., 2017; Matoušková *et al*., 2006). Cette observation, qui s'étend aussi à d'autres transposons, constitue une des caractéristiques connues du méthylome placentaire (MacAulay *et al*., 2011; Reiss *et al*., 2007; Robinson et Price, 2015). Certains résultats en lien avec les tDMR, mais aussi avec les mQTL, trouveraient peut-être une explication biologique s'ils étaient réexaminés sous cet angle, particulièrement en ce qui concerne les résultats dans des régions qui ne sont pas associées à un gène. S'il s'avérait que des mQTL étaient associés à la mADN des transposons (et potentiellement à leur répression transcriptionnelle), il serait intéressant de se questionner sur leurs implications potentielles dans le développement de cancer puisque l'activation dans des cellules cancéreuses de transposons dont l'expression est propre au placenta pourrait jouer un rôle dans leur capacité d'invasion cellulaire (Macaulay *et al*., 2017).

4.2 Limites

Afin de bien délimiter les retombées du travail qui a été présenté, il est important d'en comprendre les limites. D'abord, la composition de la cohorte Gen3G est principalement d'ascendance européenne. Pour les analyses présentées dans ce mémoire, le faible nombre d'individus dont l'ascendance était différente n'était pas suffisant pour qu'ils soient inclus. Il y a ainsi un défaut de représentativité de la diversité génétique de la population à l'étude, ce qui s'inscrit malgré nous dans le contexte connu et décrié de la surreprésentation des populations

d'ascendance européennes en science (Popejoy et Fullerton, 2016). Ensuite arrivent les limitations plus évidentes et sans doute universelles, soit celles liées à la taille des données. Une plus grande cohorte et des recoupements considérables entre les jeux de données conféreraient plus de puissance statistique aux analyses, permettant de détecter de plus petites tailles d'effet et de d'inclure potentiellement plus d'individus d'ascendance non européenne. Cependant, étant donné la taille respectable de la cohorte Gen3G, ces aspects ne sont probablement pas les plus limitants au niveau des analyses et de l'interprétation des données. Les informations de composition cellulaire et de phasage des données génomiques, en revanche, seraient des atouts précieux pour bonifier les analyses. Bien que la composition cellulaire de nos échantillons de placenta et de sang de cordon n'ait pas été déterminée en laboratoire, il est possible de l'estimer grâce à la mADN (Aryee *et al*., 2014; Fortin *et al*., 2017; Yuan *et al*., 2021). Ces estimations minimisent les impacts de la variation entre les échantillons dans nos analyses et compensent de manière suffisante l'absence de l'information mesurée (Jaffe et Irizarry, 2014). Cependant, aucune solution équivalente n'existe pour le phasage des données génomiques et de mADN. S'il était possible d'avoir cette information, elle permettrait de mener de réels tests d'association proximale entre les SNP et les CpG, puisqu'il serait possible d'intégrer entre elles des informations provenant de la même copie du chromosome et sans devoir présumer d'un effet additif du génotype sur les niveaux de mADN. De plus, avec l'information de phasage, il serait possible de prévoir un traitement à part pour les régions sous empreinte parentale. Ceci permettrait d'une part de se questionner sur l'impact du génotype sur de grandes régions sous fort contrôle épigénétique répressif et établis tôt dans le développement (Monk *et al*., 2019). D'autre part, il deviendrait alors possible d'identifier des mQTL de manière classique dans l'autre moitié des données sans que la mADN des copies sous empreinte ne vienne modérer ces potentielles associations.

4.3 Travaux futurs

Comme il a pu être vu à la section 3.3.7, il existe une certaine complémentarité ou continuité entre les tDMR et les mQTL. Alors que dans le premier cas on s'intéresse à des régions où l'ADN est différemment méthylée entre deux tissus, on identifie dans le second cas les

déterminants génétiques des niveaux de mADN. L'exemple du gène de *PDE4B* montre que les CpG d'un même tDMR sont, lorsque impliqués dans une association mQTL, généralement associés au même SNP. Cette observation s'appliquait pour 323 autres groupes de CpG dans des tDMR (section 3.3.4). Le caractère pléiotropique des SNP impliqués dans mQTL a d'ailleurs déjà été rapporté dans plusieurs tissus (Oliva *et al*., 2022). Ceci suggère que plusieurs tDMR capturent des régions où la mADN n'est pas sous contrôle du même locus génétique entre le placenta et le sang de cordon. Les résultats présentés dans ce mémoire ouvrent donc la porte à de nombreuses suites, autant pour en augmenter la qualité que pour investiguer des questions qu'ils suscitent.

D'abord, comme il a déjà été montré que la localisation des QTL présente des enrichissements pour les sites de liaison à l'ADN de plusieurs facteurs de transcription (Delahaye *et al*., 2018; Do *et al*., 2017), il serait intéressant d'également faire cette analyse pour les mQTL de placenta, d'autant plus que la mADN dans le placenta est particulièrement différente de celle des autres tissus. D'ailleurs, une des forces de la cohorte Gen3G est d'avoir des échantillons biologiques de placenta et de sang de cordon pour les mêmes individus, en plus d'échantillons de sang à cinq ans. L'identification des mQTL dans ces deux autres jeux de données sera indubitablement menée dans le futur, ce qui permettra non seulement d'évaluer la stabilité des mQTL de sang dans le temps (Gaunt *et al*., 2016), mais aussi de comparer deux ensembles de mQTL de tissus récoltés à la naissance. Cette analyse permettra de distinguer des mQTL affectant des patrons épigénétiques partagés entre les tissus somatiques de ceux affectant, par exemple, des gènes propres à un tissu (A. K. Smith *et al*., 2014), surtout si ces résultats sont mis en correspondance avec ceux de Oliva *et al.* (2022). Ces futurs résultats seront particulièrement pertinents pour mieux comprendre le méthylome placentaire. Sachant que plus de 70% des sites CpG mesurés par la puce 850K sont différemment méthylés entre le placenta et le sang de cordon, il sera intéressant de vérifier s'il existe une relation entre le petit sous-ensemble de sites similairement méthylé et les mQTL. Une fois ces analyses faites, les tDMR pourraient être réinterprétés à la lumière des mQTL qui différeront entre le placenta et le sang de cordon ombilical, puisqu'on a vu que certaines régions tDMR semblaient associées à un SNP.

Ensuite, considérant le dernier point abordé et le fait que la mADN des CpG adjacents est souvent corrélée (Eckhardt *et al*., 2006), les analyses d'identification de mQTL pourraient être refaites en tirant profit de ces informations pour identifier les SNP associés à un locus de mADN. La plus récente version de *TensorQTL*, maintenant intégrée à une suite nommée *QTLTools*, offre deux méthodes pour identifier les QTL associés à des groupes de phénotypes, lesquelles tiennent compte de tous les phénotypes dans une groupe donné (en l'occurrence, des CpG) ainsi que de la structure de corrélation entre eux. Les résultats de mQTL ainsi obtenus prennent en considération les tests multiples liés aux SNPs, aux CpG ainsi que la corrélation entre les phénotypes. L'utilisation de cette méthode dans des travaux futurs permettrait de contourner de potentiels biais introduits par la corrélation entre les CpG et d'obtenir une couche d'information supplémentaire quant à la régulation de la mADN à l'échelle d'une région génomique, et non seulement individuelle.

Finalement, comme des eQTL ont été produits dans le placenta, leur colocalisation avec les mQTL pourra être évaluée, même si les récents travaux de Oliva *et al.* (2022) laissent penser que la fraction colocalisée de mQTL et de eQTL sera faible. Ceci ne limite toutefois pas le potentiel informatif de cette analyse, surtout si certains de ces mQTL ou eQTL s'avèrent être spécifiques au placenta.

En somme, ces travaux nous rappellent la complexité de l'épigénome et les multiples angles sous lesquels il peut être étudié, même quand on ne s'intéresse qu'à une marque spécifique, dans un seul tissu. À cette complexité viennent s'ajouter de nouveau défis. Par exemple, plusieurs revues de littérature récentes soulignent que, malgré la présence d'études contradictoires, les technologies de reproduction assistées semblent être liées à des modification épigénétiques du placenta, ce qui pourrait avoir une incidence sur la santé de ces enfants (Mani *et al*., 2020; Rhon-Calderon *et al*., 2019; Sundrani et Joshi, 2021). De même, il semble devenir évident que l'exposition prénatale aux polluants affecte elle aussi l'épigénome placentaire (Ghazi *et al*., 2021; Isaevska *et al*., 2021; Lapehn et Paquette, 2022). Dans une perspective de DOHaD, ces observations pourraient être lourdes d'implications, et une bonne compréhension de la mADN

durant le développement est nécessaire pour les interpréter et en évaluer les conséquences. De nombreuses et importantes pistes restent donc à être explorées pour mieux comprendre le méthylome placentaire, sa régulation et le rôle qu'il a à jouer dans le développement et la santé à long terme de l'enfant.

ANNEXE

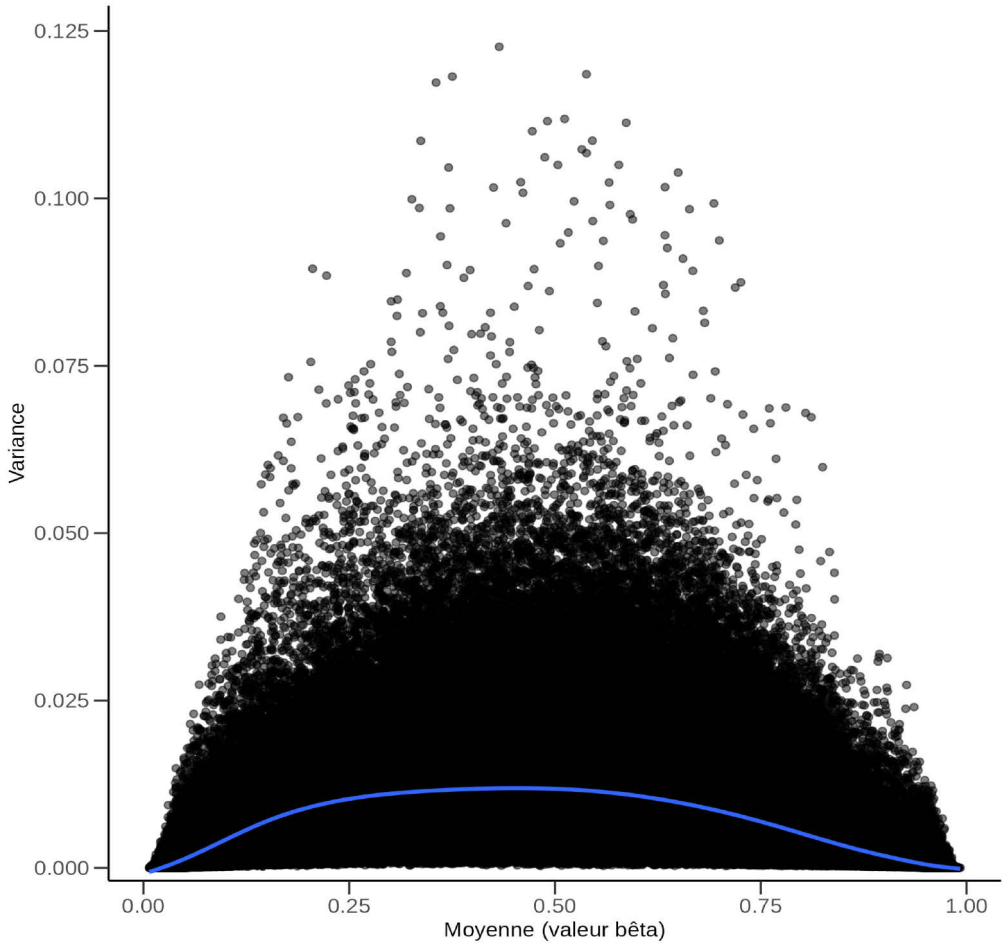

Figure S A. 1 La variance du niveau de mADN est en moyenne plus élevée pour les sites CpG dont les valeurs de méthylation sont intermédiaires.

Chaque point (N = 681 795) représente un CpG selon sa moyenne et sa variance à travers tous les individus du jeu de données (N = 524). La ligne en bleu représente la tendance de la relation entre la moyenne de mADN et sa variance, et a été générée avec les valeurs par défaut de la fonction *geom_smooth()* de la librairie R *ggplot2*.

BIBLIOGRAPHIE

Abdellah, Z., Ahmadi, A., Ahmed, S., Aimable, M., Ainscough, R., Almeida, J., Almond, C., Ambler, A., Ambrose, K., Ambrose, K., *et al*. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Adsera, C. B., Park, Y. P., Meuleman, W., et Kellis, M. (2019). Integrative analysis of 10,000 epigenomic maps across 800 samples for regulatory genomics and disease dissection. *BioRxiv*, 810291. https://doi.org/10.1101/810291

Aguet, F., Barbeira, A. N., Bonazzola, R., Brown, A., Castel, S. E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., *et al*. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, *369*(6509), 1318–1330. https://doi.org/10.1126/science.aaz1776

Alexa, A., Rahnenführer, J., et Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, *22*(13), 1600–1607. https://doi.org/10.1093/BIOINFORMATICS/BTL140

Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S. B., *et al*. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, *491*(7422), 56–65. https://doi.org/10.1038/nature11632

Aryee, M. J., Jaffe, A. E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A. P., Hansen, K. D., et Irizarry, R. A. (2014). Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, *30*(10), 1363–1369. https://doi.org/10.1093/bioinformatics/btu049

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Bentley, D. R., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., *et al*. (2015). A global reference for human genetic variation. In *Nature* (Vol. 526, Issue 7571, pp. 68–74). Nature Publishing Group. https://doi.org/10.1038/nature15393

Ball, M. P., Li, J. B., Gao, Y., Lee, J.-H. H., LeProust, E. M., Park, I.-H. H., Xie, B., Daley, G. Q., et Church, G. M. (2009). Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature Biotechnology*, *27*(4), 361–368. https://doi.org/10.1038/nbt.1533

Barker, D. J. P. (1988). Childhood causes of adult diseases. In *Archives of Disease in Childhood* (Vol. 63).

Beaumont, R. N., Flatley, C., Vaudel, M., Wu, X., Chen, J., Moen, G.-H., Skotte, L., Helgeland, Ø., Sole-Navais, P., Banasik, K., *et al*. (2022). Genome-wide association study of placental weight in 179,025 children and parents reveals distinct and shared genetic influences between placental and fetal growth. In *medRxiv*. https://doi.org/10.1101/2022.11.25.22282723

Bianco-Miotto, T., Mayne, B. T., Buckberry, S., Breen, J., Rodriguez Lopez, C. M., et Roberts, C. T. (2016). Recent progress towards understanding the role of DNA methylation in human placental development. *Reproduction (Cambridge, England)*, *152*(1), R23-30. https://doi.org/10.1530/REP-16-0014

Bolze, P. A., Mommert, M., et Mallet, F. (2017). Contribution of Syncytins and Other Endogenous Retroviral Envelopes to Human Placenta Pathologies. *Progress in Molecular Biology and Translational Science*, *145*, 111–162. https://doi.org/10.1016/BS.PMBTS.2016.12.005

Bourc'his, D., et Bestor, T. H. (2004). Meiotic catastrophe and retrotransposon reactivation in male germ cells lacking Dnmt3L. *Nature*, *431*(7004), 96–99. https://doi.org/10.1038/nature02886

Breton, C. V., Marsit, C. J., Faustman, E., Nadeau, K., Goodrich, J. M., Dolinoy, D. C., Herbstman, J., Holland, N., LaSalle, J. M., Schmidt, R., *et al*. (2017). Small-Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. *Environmental Health Perspectives*, *125*(4), 511. https://doi.org/10.1289/EHP595

Broad Institute. (n.d.). *Analysis pipelines for the GTEx Consortium and TOPMed*. https://github.com/broadinstitute/gtex-pipeline

Bujold, D., Morais, D. A. de L., Gauthier, C., Côté, C., Caron, M., Kwan, T., Chen, K. C., Laperle, J., Markovits, A. N., Pastinen, T., Caron, B., Veilleux, A., Jacques, P. É., et Bourque, G. (2016). The International Human Epigenome Consortium Data Portal. *Cell Systems*, *3*(5), 496-499.e2. https://doi.org/10.1016/j.cels.2016.10.019

Cardenas, A., Allard, C., Doyon, M., Houseman, E. A., Bakulski, K. M., Perron, P., Bouchard, L., et Hivert, M. F. (2016). Validation of a DNA methylation reference panel for the estimation of nucleated cells types in cord blood. *Epigenetics*, *11*(11), 773–779. https://doi.org/10.1080/15592294.2016.1233091

Cardenas, A., Gagné-Ouellet, V., Allard, C., Brisson, D., Perron, P., Bouchard, L., et Hivert, M. F. (2018). Placental DNA Methylation Adaptation to Maternal Glycemic Response in Pregnancy. *Diabetes*, *67*(8), 1673–1683. https://doi.org/10.2337/db18-0123

Cavalli, G., et Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature*, *571*(7766), 489–499. https://doi.org/10.1038/s41586-019-1411-0

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., et Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), 7. https://doi.org/10.1186/s13742-015-0047-8

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H. C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., *et al*. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*(7). https://doi.org/10.1371/journal.pbio.1001091

Conomos, M. P., Miller, M. B., et Thornton, T. A. (2015). Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology*, *39*(4), 276–293. https://doi.org/10.1002/gepi.21896

Conomos, M. P., Reiner, A. P., Weir, B. S., et Thornton, T. A. (2016). Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics*, *98*(1), 127–148. https://doi.org/10.1016/j.ajhg.2015.11.022

Cooper, D. N., et Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. *Human Genetics*, *83*(2), 181–188. https://doi.org/10.1007/BF00286715

Costa, M. A. (2016). The endocrine function of human placenta: an overview. *Reproductive BioMedicine Online*, *32*(1), 14–43. https://doi.org/10.1016/J.RBMO.2015.10.005

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, *573*(7772), 45–54. https://doi.org/10.1038/s41586-019-1517-4

Das, S., Forer, L., Schönherr, S., Sidore, C., Locke, A. E., Kwong, A., Vrieze, S. I., Chew, E. Y., Levy, S., McGue, M., *et al*. (2016). Next-generation genotype imputation service and methods. *Nature Genetics*, *48*(10), 1284–1287. https://doi.org/10.1038/ng.3656

Deaton, A. M., et Bird, A. (2011). CpG islands and the regulation of transcription. *Genes* et *Development*, *25*(10), 1010–1022. https://doi.org/10.1101/gad.2037511

Delahaye, F., Do, C., Kong, Y., Ashkar, R., Salas, M., Tycko, B., Wapner, R., et Hughes, F. (2018). Genetic variants influence on the placenta regulatory landscape. In *PLoS Genetics* (Vol. 14, Issue 11). https://doi.org/10.1371/journal.pgen.1007785

Delaneau, O., Ongen, H., Brown, A. A., Fort, A., Panousis, N. I., et Dermitzakis, E. T. (2017). A complete tool set for molecular QTL discovery and analysis. *Nature Communications*, *8*(1), 15452. https://doi.org/10.1038/ncomms15452

Deluca, D. S., Levin, J. Z., Sivachenko, A., Fennell, T., Nazaire, M. D., Williams, C., Reich, M., Winckler, W., et Getz, G. (2012). RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*, *28*(11), 1530–1532. https://doi.org/10.1093/bioinformatics/bts196

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V, Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., *et al*. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), 491–498. https://doi.org/10.1038/ng.806

Dergai, O., et Hernandez, N. (2019). How to Recruit the Correct RNA Polymerase? Lessons from snRNA Genes. *Trends in Genetics*, *35*(6), 457–469. https://doi.org/10.1016/j.tig.2019.04.001

Do, C., Shearer, A., Suzuki, M., Terry, M. B., Gelernter, J., Greally, J. M., et Tycko, B. (2017). Genetic-epigenetic interactions in cis: A major focus in the post-GWAS era. *Genome Biology*, *18*(1), 1–22. https://doi.org/10.1186/s13059-017-1250-y

Du, P., Zhang, X., Huang, C. C., Jafari, N., Kibbe, W. A., Hou, L., et Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics*, *11*(1), 587. https://doi.org/10.1186/1471-2105-11-587

Du, Q., Luu, P. L., Stirzaker, C., et Clark, S. J. (2015). Methyl-CpG-binding domain proteins: Readers of the epigenome. In *Epigenomics* (Vol. 7, Issue 6, pp. 1051–1073). Future Medicine Ltd. https://doi.org/10.2217/epi.15.39

Eckhardt, F., Lewin, J., Cortese, R., Rakyan, V. K., Attwood, J., Burger, M., Burton, J., Cox, T. V, Davies, R., Down, T. A., *et al*. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nature Genetics*, *38*(12), 1378–1385. https://doi.org/10.1038/ng1909

Edwards, J. R., Yarychkivska, O., Boulard, M., et Bestor, T. H. (2017). DNA methylation and DNA methyltransferases. *Epigenetics* et *Chromatin*, *10*(1), 23. https://doi.org/10.1186/s13072-017-0130-8

Epi-centered Research. (n.d.). *QC Steps for PACE Analyses*. Retrieved July 21, 2023, from https://www.epicenteredresearch.com/pace/qcsteps/

Ernst, J., et Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, *28*(8), 817–825. https://doi.org/10.1038/nbt.1662

Felix, J. F., Joubert, B. R., Baccarelli, A. A., Sharp, G. C., Almqvist, C., Annesi-Maesano, I., Arshad, H., Baiz, N., Bakermans-Kranenburg, M. J., Bakulski, K. M., *et al*. (2018). Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *International Journal of Epidemiology*, *47*(1), 22–23u. https://doi.org/10.1093/IJE/DYX190

Felsenfeld, G., et Groudine, M. (2003). Controlling the double helix. *Nature*, *421*(6921), 448–453. https://doi.org/10.1038/nature01411

Fields, C., Adams, M. D., White, O., et Craig Venter, J. (1994). *How many genes in the human genome?*

Fortin, J. P., Labbé, A., Zanke, B. W., Hudson, T. J., Fertig, E. J., Greenwood, C. M. T., et Hansen, K. D. (2014). Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology*, *98*(4), 288–295.

Fortin, J. P., Triche, T. J., et Hansen, K. D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics*, *33*(4), 558–560. https://doi.org/10.1093/bioinformatics/btw691

Frankish, A., Diekhans, M., Ferreira, A. M., Johnson, R., Jungreis, I., Loveland, J., Mudge, J. M., Sisu, C., Wright, J., Armstrong, J., *et al*. (2019). GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Research*, *47*(D1), D766–D773. https://doi.org/10.1093/NAR/GKY955

Gaunt, T. R., Shihab, H. A., Hemani, G., Min, J. L., Woodward, G., Lyttleton, O., Zheng, J., Duggirala, A., McArdle, W. L., Ho, K., *et al*. (2016). Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, *17*, 61. https://doi.org/10.1186/s13059-016-0926-z

Gendrel, A. V., Apedaile, A., Coker, H., Termanis, A., Zvetkova, I., Godwin, J., Tang, Y. A., Huntley, D., Montana, G., Taylor, S., *et al*. (2012). Smchd1-Dependent and -Independent Pathways Determine Developmental Dynamics of CpG Island Methylation on the Inactive X Chromosome. *Developmental Cell*, *23*(2), 265–279. https://doi.org/10.1016/j.devcel.2012.06.011

Ghazi, T., Naidoo, P., Naidoo, R. N., et Chuturgoon, A. A. (2021). Prenatal Air Pollution Exposure and Placental DNA Methylation Changes: Implications on Fetal Development and Future Disease Susceptibility. *Cells*, *10*(11). https://doi.org/10.3390/CELLS10113025

Gimpl, G., et Fahrenholz, F. (2001). The Oxytocin Receptor System: Structure, Function, and Regulation. *Physiological Reviews*, *81*(2), 629–683. https://doi.org/10.1152/physrev.2001.81.2.629

Gogarten, S. M., Sofer, T., Chen, H., Yu, C., Brody, J. A., Thornton, T. A., Rice, K. M., et Conomos, M. P. (2019). Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btz567

g:Profiler. (n.d.). *g:GOSt documentation on highlighting*. Retrieved July 21, 2023, from https://biit.cs.ut.ee/gprofiler/page/docs#highlighting-description

Graubert, A., Aguet, F., Ravi, A., Ardlie, K. G., et Getz, G. (2021). RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts. *Bioinformatics*, *37*(18), 3048–3050. https://doi.org/10.1093/BIOINFORMATICS/BTAB135

Greenberg, M. V. C., et Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology*. https://doi.org/10.1038/s41580-019-0159-6

Groleau, M., White, F., Cardenas, A., Perron, P., Hivert, M.-F., Bouchard, L., et Jacques, P.-É. (2021). Comparative epigenome-wide analysis highlights placenta-specific differentially methylated regions. *Epigenomics*, *13*(5), 357–368. https://doi.org/10.2217/epi-2020-0271

Guillemette, L., Allard, C., Lacroix, M., Patenaude, J., Battista, M. C., Doyon, M., Moreau, J., Ménard, J., Bouchard, L., Ardilouze, J. L., Perron, P., et Hivert, M. F. (2016). Genetics of Glucose regulation in Gestation and Growth (Gen3G): A prospective prebirth cohort of mother-child pairs in Sherbrooke, Canada. *BMJ Open*, *6*(2), 1–14. https://doi.org/10.1136/bmjopen-2015-010031

Heather, J. M., et Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. In *Genomics* (Vol. 107, Issue 1, pp. 1–8). Academic Press Inc. https://doi.org/10.1016/j.ygeno.2015.11.003

Heiss, J. A., et Just, A. C. (2018). Identifying mislabeled and contaminated DNA methylation microarray data: An extended quality control toolset with examples from GEO. *Clinical Epigenetics*, *10*(1), 1–9. https://doi.org/10.1186/S13148-018-0504-1/FIGURES/5

Henikoff, S., et Smith, M. M. (2015). Histone variants and epigenetics. *Cold Spring Harbor Perspectives in Biology*, *7*(1), 1–25. https://doi.org/10.1101/cshperspect.a019364

Hoffman, D. J., Reynolds, R. M., et Hardy, D. B. (2017). Developmental origins of health and disease: Current knowledge and potential mechanisms. *Nutrition Reviews*, *75*(12), 951–970. https://doi.org/10.1093/nutrit/nux053

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, *14*(10), R115. https://doi.org/10.1186/gb-2013-14-10-r115

Hrdlickova, R., Toloue, M., et Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, *8*(1), e1364. https://doi.org/10.1002/wrna.1364

Hyun, K., Jeon, J., Park, K., et Kim, J. (2017). Writing, erasing and reading histone lysine methylations. In *Experimental and Molecular Medicine* (Vol. 49, Issue 4, p. 324). Nature Publishing Group. https://doi.org/10.1038/emm.2017.11

Isaevska, E., Moccia, C., Asta, F., Cibella, F., Gagliardi, L., Ronfani, L., Rusconi, F., Stazi, M. A., et Richiardi, L. (2021). Exposure to ambient air pollution in the first 1000 days of life and alterations in the DNA methylome and telomere length in children: A systematic review. *Environmental Research*, *193*. https://doi.org/10.1016/J.ENVRES.2020.110504

Jaffe, A. E., et Irizarry, R. A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, *15*(R31), 1–9. https://doi.org/10.1186/gb-2014-15-2-r31

Johnson, W. E., Li, C., et Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, *8*(1), 118–127. https://doi.org/10.1093/biostatistics/kxj037

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., *et al*. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kerimov, N., Hayhurst, J., Manning, J., Walter, P., Kolberg, L., Peikova, K., Samoviča, M., Burdett, T., Jupp, S., Parkinson, H., Papatheodorou, I., Zerbino, D., et Alasoo, K. (2020). eQTL Catalogue: a compendium of uniformly processed human gene expression and splicing QTLs. *BioRxiv*, 2020.01.29.924266. https://doi.org/10.1101/2020.01.29.924266

Kohli, R. M., et Zhang, Y. (2013). TET enzymes, TDG and the dynamics of DNA demethylation. In *Nature* (Vol. 502, Issue 7472, pp. 472–479). Nature Publishing Group. https://doi.org/10.1038/nature12750

Koukoura, O., Sifakis, S., et Spandidos, D. A. (2012). DNA methylation in the human placenta and fetal growth (review). *Molecular Medicine Reports*, *5*(4), 883–889. https://doi.org/10.3892/mmr.2012.763

Lapehn, S., et Paquette, A. G. (2022). The Placental Epigenome as a Molecular Link Between Prenatal Exposures and Fetal Health Outcomes Through the DOHaD Hypothesis. *Current Environmental Health Reports*, *9*(3), 490–501. https://doi.org/10.1007/S40572-022-00354-8

Li, E., et Zhang, Y. (2014). DNA Methylation in Mammals. *Cold Spring Harbor Perspectives in Biology*, *6*(5), a019133–a019133. https://doi.org/10.1101/cshperspect.a019133

Liao, W.-W., Asri, M., Ebler, J., Doerr, D., Haukness, M., Hickey, G., Lu, S., Lucas, J. K., Monlong, J., Abel, H. J., *et al*. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Long, H. K., Sims, D., Heger, A., Blackledge, N. P., Kutter, C., Wright, M. L., Grützner, F., Odom, D. T., Patient, R., Ponting, C. P., et Klose, R. J. (2013). Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates. *ELife*, *2*(2). https://doi.org/10.7554/eLife.00348

Macaulay, E. C., Chatterjee, A., Cheng, X., Baguley, B. C., Eccles, M. R., et Morison, I. M. (2017). The Genes of Life and Death: A Potential Role for Placental-Specific Genes in Cancer. *BioEssays*, *39*(11), 1700091. https://doi.org/10.1002/BIES.201700091

MacAulay, E. C., Weeks, R. J., Andrews, S., et Morison, I. M. (2011). Hypomethylation of functional retrotransposon-derived genes in the human placenta. *Mammalian Genome*, *22*(11–12), 722–735. https://doi.org/10.1007/S00335-011-9355-1/FIGURES/8

Maltepe, E., et Fisher, S. J. (2015). Placenta: The Forgotten Organ. *Annual Review of Cell and Developmental Biology*, *31*(1), 523–552. https://doi.org/10.1146/annurev-cellbio-100814-125620

Mani, S., Ghosh, J., Coutifaris, C., Sapienza, C., et Mainigi, M. (2020). Epigenetic changes and assisted reproductive technologies. *Epigenetics*, *15*(1–2), 12. https://doi.org/10.1080/15592294.2019.1646572

Marasca, F., Bodega, B., et Orlando, V. (2018). How Polycomb-Mediated Cell Memory Deals With a Changing Environment. *BioEssays*, *40*(4), 1700137. https://doi.org/10.1002/bies.201700137

Matoušková, M., Blažková, J., Pajer, P., Pavlíček, A., et Hejnar, J. (2006). CpG methylation suppresses transcriptional activity of human syncytin-1 in non-placental tissues. *Experimental Cell Research*, *312*(7), 1011–1020. https://doi.org/10.1016/J.YEXCR.2005.12.010

Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R., et Riccio, A. (2019). Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics*, *20*(4), 235–248. https://doi.org/10.1038/s41576-018-0092-0

Moore, J. E., Purcaro, M. J., Pratt, H. E., Epstein, C. B., Shoresh, N., Adrian, J., Kawli, T., Davis, C. A., Dobin, A., Kaul, R., *et al*. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, *583*(7818), 699–710. https://doi.org/10.1038/s41586-020-2493-4

NIH. (2020, August 16). *Epigenomics Fact Sheet*. https://www.genome.gov/about-genomics/fact-sheets/Epigenomics-Fact-Sheet

NIH. (2021, June 11). *What is epigenetics?* https://medlineplus.gov/genetics/understanding/howgeneswork/epigenome/

Okano, M., Bell, D. W., Haber, D. A., et Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell*, *99*(3), 247–257. https://doi.org/10.1016/S0092-8674(00)81656-6

Oliva, M., Demanelis, K., Lu, Y., Chernoff, M., Jasmine, F., Ahsan, H., Kibriya, M. G., Chen, L. S., et Pierce, B. L. (2022). DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature Genetics 2022 55:1*, *55*(1), 112–122. https://doi.org/10.1038/s41588-022-01248-z

Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T., et Delaneau, O. (2016). Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, *32*(10), 1479–1485. https://doi.org/10.1093/bioinformatics/btv722

Orsi, G. A., Couble, P., et Loppin, B. (2009). Epigenetic and replacement roles of histone variant H3.3 in reproduction and development. *The International Journal of Developmental Biology*, *53*(2–3), 231–243. https://doi.org/10.1387/ijdb.082653go

Ozanne, S. E., et Constância, M. (2007). Mechanisms of Disease: The developmental origins of disease and the role of the epigenotype. *Nature Clinical Practice Endocrinology and Metabolism*, *3*(7), 539–546. https://doi.org/10.1038/ncpendmet0531

Patil, V., Ward, R. L., et Hesson, L. B. (2014). The evidence for functional non-CpG methylation in mammalian cells. In *Epigenetics* (Vol. 9, Issue 6, pp. 823–828). Taylor and Francis Inc. https://doi.org/10.4161/epi.28741

Pidsley, R., Wong, C. C. Y., Volta, M., Lunnon, K., Mill, J., et Schalkwyk, L. C. (2013). A data-driven approach to preprocessing Illumina 450 K methylation array data. *BMC Genomics*, *14*, 293. https://doi.org/10.1186/1471-2164-14-293

Pingault, J.-B., O'Reilly, P. F., Schoeler, T., Ploubidis, G. B., Rijsdijk, F., et Dudbridge, F. (2018). Using genetic data to strengthen causal inference in observational research. *Nature Reviews Genetics*, *19*(9), 566–580. https://doi.org/10.1038/s41576-018-0020-3

Piovesan, A., Antonaros, F., Vitale, L., Strippoli, P., Pelleri, M. C., et Caracausi, M. (2019). Human protein-coding genes and gene feature statistics in 2019. *BMC Research Notes*, *12*(1), 315. https://doi.org/10.1186/s13104-019-4343-8

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., et Sandhu, M. S. (2018). Long reads: their purpose and place. In *Human molecular genetics* (Vol. 27, Issue R2, pp. R234–R241). NLM (Medline). https://doi.org/10.1093/hmg/ddy177

Popejoy, A. B., et Fullerton, S. M. (2016). Genomics is failing on diversity. *Nature*, *538*(7624), 161–164. https://doi.org/10.1038/538161a

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Auwera, G. A. Van der, Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., *et al*. (2017). Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, 201178. https://doi.org/10.1101/201178

Quinlan, A. R., et Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/BIOINFORMATICS/BTQ033

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., et Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, *47*(W1), W191–W198. https://doi.org/10.1093/nar/gkz369

Reiss, D., Zhang, Y., et Mager, D. L. (2007). Widely variable endogenous retroviral methylation levels in human placenta. *Nucleic Acids Research*, *35*(14), 4743. https://doi.org/10.1093/NAR/GKM455

Rhon-Calderon, E. A., Vrooman, L. A., Riesche, L., et Bartolomei, M. S. (2019). The effects of Assisted Reproductive Technologies on genomic imprinting in the placenta. *Placenta*, *84*, 37–43. https://doi.org/10.1016/J.PLACENTA.2019.02.013

Robinson, W. P., et Price, E. M. (2015). The human placental methylome. *Cold Spring Harbor Perspectives in Medicine*, *5*(5), 1–15. https://doi.org/10.1101/cshperspect.a023044

Roeder, R. G., et Rutter, W. J. (1969). Multiple Forms of DNA-dependent RNA Polymerase in Eukaryotic Organisms. *Nature*, *224*(5216), 234–237. https://doi.org/10.1038/224234a0

Rondinone, O., Murgia, A., Costanza, J., Tabano, S., Camanni, M., Corsaro, L., Fontana, L., Colapietro, P., Calzari, L., Motta, S., *et al*. (2021). Extensive Placental Methylation Profiling in Normal Pregnancies. *International Journal of Molecular Sciences 2021, Vol. 22, Page 2136*, *22*(4), 2136. https://doi.org/10.3390/IJMS22042136

Schneider, V. A., Graves-Lindsay, T., Howe, K., Bouk, N., Chen, H. C., Kitts, P. A., Murphy, T. D., Pruitt, K. D., Thibaud-Nissen, F., Albracht, D., *et al*. (2017). Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, *27*(5), 849–864. https://doi.org/10.1101/gr.213611.116

Schroeder, D. I., Blair, J. D., Lott, P., Yu, H. O. K., Hong, D., Crary, F., Ashwood, P., Walker, C., Korf, I., Robinson, W. P., et LaSalle, J. M. (2013). The human placenta methylome. *Proceedings of the National Academy of Sciences*, *110*(15), 6037–6042. https://doi.org/10.1073/pnas.1215145110

Schulz, L. C. (2010). The Dutch Hunger Winter and the developmental origins of health and disease. *Proceedings of the National Academy of Sciences*, *107*(39), 16757–16758. https://doi.org/10.1073/pnas.1012911107

Sharif, J., Muto, M., Takebayashi, S. I., Suetake, I., Iwamatsu, A., Endo, T. A., Shinga, J., Mizutani-Koseki, Y., Toyoda, T., Okamura, K., *et al*. (2007). The SRA protein Np95 mediates epigenetic inheritance by recruiting Dnmt1 to methylated DNA. *Nature*, *450*(7171), 908–912. https://doi.org/10.1038/nature06397

Shayevitch, R., Askayo, D., Keydar, I., et Ast, G. (2018). The importance of DNA methylation of exons on alternative splicing. *RNA*, *24*(10), 1351–1362. https://doi.org/10.1261/rna.064865.117

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., et Darvasi, A. (2003). Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics*, *12*(7), 771–776. https://doi.org/10.1093/HMG/DDG088

Smith, A. K., Kilaru, V., Kocak, M., Almli, L. M., Mercer, K. B., Ressler, K. J., Tylavsky, F. A., et Conneely, K. N. (2014). Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, *15*(1), 145. https://doi.org/10.1186/1471-2164-15-145

Smith, G. D., et Hemani, G. (2014). Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, *23*(R1), 89–98. https://doi.org/10.1093/hmg/ddu328

Stegle, O., Parts, L., Piipari, M., Winn, J., et Durbin, R. (2012). Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, *7*(3), 500. https://doi.org/10.1038/NPROT.2011.457

Storey, J. D., et Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(16), 9440–9445. https://doi.org/10.1073/PNAS.1530509100/ASSET/25537429-365C-4D06-977A-86C871368513/ASSETS/GRAPHIC/PQ1530509003.JPEG

Sundrani, D. P., et Joshi, S. R. (2021). Assisted reproductive technology (ART) and epigenetic modifications in the placenta. *Human Fertility*. https://doi.org/10.1080/14647273.2021.1995901

Taylor-Weiner, A., Aguet, F., Haradhvala, N. J., Gosai, S., Anand, S., Kim, J., Ardlie, K., Van Allen, E. M., et Getz, G. (2019). Scaling computational genomics to millions of individuals with GPUs. *Genome Biology 2019 20:1*, *20*(1), 1–5. https://doi.org/10.1186/S13059-019-1836-7

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., et Beck, S. (2013). A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, *29*(2), 189–196. https://doi.org/10.1093/BIOINFORMATICS/BTS680

Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W., et Siegmund, K. D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research*, *41*(7), e90–e90. https://doi.org/10.1093/NAR/GKT090

Voss, A. K., et Thomas, T. (2018). Histone Lysine and Genomic Targets of Histone Acetyltransferases in Mammals. *BioEssays*, *40*(10), 1800078. https://doi.org/10.1002/bies.201800078

WhatIsEpigenetics. (2013a, July 20). *DNA Methylation*. https://www.whatisepigenetics.com/dna-methylation/

WhatIsEpigenetics. (2013b, July 20). *Histone Modifications*. https://www.whatisepigenetics.com/histone-modifications/

Xiao, C.-L., Zhu, S., He, M., Chen, D., Zhang, Q., Chen, Y., Yu, G., Liu, J., Xie, S.-Q., Luo, F., *et al*. (2018). N6-Methyladenine DNA Modification in the Human Genome. *Molecular Cell*, *71*(2), 306-318.e7. https://doi.org/10.1016/j.molcel.2018.06.015

Xu, C., Liu, K., Lei, M., Yang, A., Li, Y., Hughes, T. R., et Min, J. (2018). DNA Sequence Recognition of Human CXXC Domains and Their Structural Determinants. *Structure*, *26*(1), 85-95.e3. https://doi.org/10.1016/j.str.2017.11.022

Ye, Y., Zhang, Z., Liu, Y., Diao, L., et Han, L. (2020). A Multi-Omics Perspective of Quantitative Trait Loci in Precision Medicine. *Trends in Genetics*, *36*(5), 318–336. https://doi.org/10.1016/j.tig.2020.01.009

Yuan, V., Hui, D., Yin, Y., Peñaherrera, M. S., Beristain, A. G., et Robinson, W. P. (2021). Cell-specific characterization of the placental methylome. *BMC Genomics*, *22*(1), 1–20. https://doi.org/10.1186/S12864-020-07186-6/TABLES/3

Zhou, V. W., Goren, A., et Bernstein, B. E. (2011). Charting histone modifications and the functional organization of mammalian genomes. In *Nature Reviews Genetics* (Vol. 12, Issue 1, pp. 7–18). Nature Publishing Group. https://doi.org/10.1038/nrg2905