Épidémiologie épi-génétique de biomarqueurs du risque cardiovasculaire : intérêt de l'étude de la méthylation de l'ADN à partir d'échantillons sanguins

Dylan Aïssi

NNT : 2015SACLS011

THÈSE DE DOCTORAT

de

l'Université Paris-Saclay

préparée à

l'Université Paris-Sud

École Doctorale n° 570 : Santé Publique

Spécialité doctorale "Génétique Statistique"

présentée et soutenue publiquement par

Dylan AÏSSI

le 12 octobre 2015

Épidémiologie épi-génétique de biomarqueurs du risque cardiovasculaire : intérêt de l'étude de la méthylation de l'ADN à partir d'échantillons sanguins

Directeur de thèse : **David-Alexandre TRÉGOUËT**

Composition du Jury :

Dr. Jean-Philippe EMPANA,	Paris - Centre de Recherche Cardiovasculaire, Paris	Président du jury
Pr. Murielle BOCHUD,	Institut Universitaire de Médecine Sociale et Préventive, Lausanne	Rapporteur
Dr. Zdenko HERCEG,	Centre International de Recherche sur le Cancer, Lyon	Rapporteur
Dr. Jörg TOST,	Centre National de Génotypage, Évry	Examinateur

Consider God's handiwork ; who can straighten what He hath made crooked ?

Considérez le travail de Dieu ; qui peut redresser ce qu'il aurait tordu ?

Ecclesiastes 7 :13

I not only think that we will tamper with Mother Nature, I think Mother wants us to.

Je ne pense pas seulement que nous allons altérer notre mère nature, je pense que c'est ce qu'elle attend de nous.

Willard Gaylin

There is no gene for the human spirit.

Il n'y a pas de gènes pour la volonté humaine.

Gattaca, 1997

Remerciements

En premier lieu, je tiens à remercier mon Directeur de thèse David-Alexandre Trégouët qui a eu confiance en moi et qui a su me guider au cours de ces trois années de thèse. Je le remercie de s'être aventuré avec moi dans l'épidémiologie épigénétique sans savoir ce qui nous attendrait. Comme il me l'a si bien dit "*on ne sait pas qui nage nu dans l'océan tant que la marée ne s'est pas retirée !*". Il a su me motiver lorsque "*j'avais du pain sur la planche et que ce n'était pas du pain de mie !*". Je le remercie pour tout ce qu'il m'a appris de scientifique et de non scientifique. Son sens de l'humour aura été "*le cornichon sur le pâté*". Merci David !

Je remercie mes rapporteurs, Madame la Professeure Murielle Bochud et Monsieur Zdenko Herceg, pour avoir lu mon manuscrit de thèse et m'avoir donné leurs avis critiques sur mon travail. Je remercie également mes examinateurs, Monsieur Jean-Philippe Empana et Monsieur Jörg Tost, d'avoir accepté de consacrer de leur temps pour être membre de mon jury.

Je remercie ensuite François Cambien et Laurence Tiret pour m'avoir accueilli au début de ma thèse au sein de l'ancienne unité INSERM UMR_S 937.

Je remercie également le Professeur Pierre-Emmanuel Morange qui a mis en place l'étude MARTHA. Je souhaite remercier France Gagnon sans qui l'étude de la méthylation dans l'étude MARTHA n'aurait pas été possible.

Je remercie le CORDDIM qui m'a permis de faire cette thèse en me finançant pendant ces trois années.

Je tiens à remercier également l'ensemble de l'équipe notamment pour la bonne ambiance qui y règne : Veronica qui, la pauvre, m'a eu comme "tutour" de français. Tout simplement, il suffit d'enlever le O ou le A à la fin des mots en italien pour avoir leurs équivalents en français ;-). Ne t'inquiète pas, ~~normalement~~ Debian fonctionnera très bien sans moi ! Henri qui a soutenu sa thèse quelques jours avant moi et qui je l'espère aura enfin trouvé son promoteur. Maguelonne alias Dame Ginette, peut-être la future chouchoute ? Je te décerne le balai d'or pour m'avoir aidé à faire le ménage dans nos nouveaux locaux au 4ème. Romain qui prend la relève en tant que Padawan de David. Beata, désolé de te laisser en galère sur le projet de protéomique. Je suis sûr que tu trouveras pourquoi les femmes sont problématiques ! Marine, l'experte des GWAS, de MARTHA et maintenant des couches-culottes. Sophie, pour sa relecture au dernier moment de mon introduction de thèse. Anne-Sophie, la dernière arrivée au labo. *Li marrons chauds li pas cheeerrr, Pakistanooooooooo*. Bon courage avec tes PC !

Varma, pour avoir partagé mon bureau dans la dernière ligne droite et supporté mes jurons sans y comprendre grand-chose. Nathalie, pour me rappeler que : "*la Bretagne, ça vous gagne !*". Et tous ceux qui ont partagé mon quotidien au cours de ces trois ans : Claire, Carole, Ewa, Hervé, etc.

Un remerciement particulier à la team de la "*one, two, three, viva l'Algérie !*" : - Zahia, pour rire à mes blagues pas souvent très drôles. - Nadjim, pour son invitation en Kabylie, ses blagues et son insistance pour avoir des RTT malgré ses deux mois de travail. - Lyamine qui nous a promis un pot de départ mais qui ne l'a jamais fait. - Badreddine, pour nos discussions sur l'amiante du bâtiment. Ne t'inquiète pas, un jour, ils enlèveront ta hôte amiantée !

Ainsi que ceux qui ont quitté le labo : Ricardo, Nicolas, Jessica, Guillaume, Vinh, Ulli , Ares et François le roux. Ainsi qu'à mon ami québécois parisien préféré Martin ! Parisien dans l'âme, mais Québécois dans sa façon de parler ! *Wo pa laille ! Tire-toi une bûche !* Un merci également à tous les Oompa Loompas stagiaires dont je ne me rappelle même plus les prénoms mais qui comme le voulait la coutume me ramenaient des croissants !

Je tiens à remercier également mes amis. Quentin avec qui j'ai fait les 400 coups dans ma jeunesse. Mon binôme Cyprien 69 et sa nouvelle binôme Marine, bientôt nous pourrons faire des combats de couches-culottes. David et Sabrina, pas de panique pour votre prochaine destination, j'ai compris comment fonctionne l'italien. Il suffit de rajouter un O ou un A à la fin des mots. Et enfin le dernier couple : Ramtinnnnnnn et Aurélien qui est parti au pays de la bière et de la saucisse !

Je remercie toute ma famille et en particulier mes grands-parents et mes parents sans qui ne je serais pas là. Ma mamie pour avoir préparé de très bonnes crêpes pour le pot de thèse. Ma maman pour avoir relu ma thèse plusieurs fois. Pas facile d'y trouver des photes lorsque cela semble être une autre langue, n'est-ce pas ? Mon papa qui m'attend tous les soirs pour prendre le RER. Ma soeurette préférée pour être... la meilleure des soeurettes qu'on puisse avoir et pour avoir fait des blagues à mes vêtements... ils étaient pliés !

Et pour finir, la plus importante, ma doudou qui réussit le double exploit de me supporter tous les jours et de créer la vie. Bien plus forte que tous les scientifiques qui ne savent toujours pas comment la vie a été créée ;-). Je te tire mon chapeau Mme Aïssi !

Articles publiés

— **DNA methylation and body-mass index : a genome-wide analysis** ; Dick K.J., Nelson C.P., Tsaprouni L., Sandling J.K., **Aïssi D.**, Wahl S., Menduri E., Morange P.E., Gagnon F., Grallert H., Waldenberger M., Peters A., Erdmann J., Hengstenberg C., Cambien F., Goodall A.H., Ouwehand W., Schunkert H., Thompson J.R., Spector T.D., Gieger C., Trégouët D.A., Deloukas P., Samani N.J. ; *The Lancet* 2014, DOI: 10.1016/S0140-6736(13)62674-4

— **Robust validation of methylation levels association at CPT1A locus with lipid plasma levels** ; Gagnon F, **Aïssi D.**, Carrié A., Morange P.E., Trégouët D.A. ; *Journal of Lipid Research* 2014, DOI: 10.1194/jlr.E051276

— **Genome-wide investigation of DNA methylation marks associated with FV Leiden mutation** ; **Aïssi D.**, Dennis J., Ladouceur M., Truong V., Zwingerman N., Rocanin-Arjo A., Germain M., Paton T.A., Morange P.E., Gagnon F., Trégouët D.A. ; *PLoS ONE* 2014, DOI: 10.1371/journal.pone.0108087

— **Thrombin generation potential and whole-blood DNA methylation** ; Rocanin-Arjo A., Dennis J., Suchon P., **Aïssi D.**, Truong V., Trégouët D.A., Gagnon F., Morange P.E. ; *Thrombosis Research* 2014, DOI: 10.1016/j.thromres.2014.12.010

— **Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci** ; Lemire M., Zaidi S.H.E., Ban M., Ge B., **Aïssi D.**, Germain M., Kassam I., Wang M., Zanke B.W, Gagnon F., Morange P.E., Trégouët D.A., Wells P.S., Sawcer S., Gallinger S., Pastinen T., Hudson T.J. ; *Nature Communications* 2015, DOI: 10.1038/ncomms7326

— **Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism** ; Germain M., Chasman D.I., de Haan H., Tang W., Lindström S., Weng L.-C., de Andrade M., de Visser M.C.H., Wiggins K.L., Suchon P., Saut N., Smadja D.M., Le Gal G., van Hylckama Vlieg A., Di Narzo A., Hao K., Nelson C.P, Rocanin-Arjo A., Folkersen L., Monajemi R., Rose L.M., Brody J.A., Slagboom E., **Aïssi D.**, Gagnon F., Deleuze J.F., Deloukas P., Tzourio C., Dartigues J.F, Berr C., Taylor K.D, Civelek M., Eriksson P., Cardiogenics Consortium, Psaty B.M., Houwing-Duitermaat J., Goo-

dall A.H., Cambien F., Kraft P., Amouyel P., Samani N.J., Basu S., Ridker P.M., Rosendaal F.R., Kabrhel C., Folsom A.R., Heit J., Reitsma P.H, Trégouët D.A., Smith N.L., Morange P.E. ; *The American Journal of Human Genetics* 2015, DOI: 10.1016/j.ajhg.2015.01.019

Communications Affichées

— **Investigation of genome-wide DNA methylation marks associated with FV Leiden mutation in patients with venous thrombosis** ; Présenté au :
  — 46ème congrès annuel de l'European Society of Human Genetics (ESHG), 8 - 11 Juin 2013, Paris, France
  — 63ème congrès annuel de l'American Society of Human Genetics (ASHG), 22 - 26 Octobre 2013, Boston, USA

— **Peripheral blood DNA is a good model for addressing the epigenetic epidemiology of cardiometabolic biomarkers** ; Présenté au :
  — 6ème congrès annuel du CORDDIM, 12 Septembre 2014, Paris, France
  — 2ème congrès annuel de l'Institute for Cardiometabolism And Nutrition (ICAN), 22 Septembre 2014, Paris, France
  — 10ème congrès annuel de l'association "Jeunes Chercheurs & Avenir - René Descartes", 8 Octobre 2014, Paris, France

Liste des principales abréviations

— **HM27k** : Puce Illumina Infinium HumanMethylation27k

— **HM450k** : Puce Illumina Infinium HumanMethylation450k

— **MARTHA** : Étude "MARseille THrombosis Association"

— **SNP** : "Single-Nucleotide Polymorphism" pour Polymorphisme Nucléotidique

— **eSNP** : SNP influençant l'expression de gènes

— **mSNP** : SNP influençant les niveaux de méthylation

— **FDR** : "False Discovery Rate" ou taux de faux positifs

— **FVL** : Facteur V Leiden

— **GWAS** : "Genome-Wide Association Study" pour étude d'association génome entier

— **GWES** : "Genome-Wide Expression Study" pour étude d'expression génome entier

— **MWAS** : "Methylome-Wide Association Study" pour étude d'association méthylome entier

— **QTL** : "Quantitative Trait Locus" pour locus de caractères quantitatifs

— **eQTL** : "expression Quantitative Trait Locus" pour locus modulant l'expression de gènes

— **eQTM** : "expression Quantitative Trait Methylation" pour niveau de méthylation modulant l'expression de gènes

— **mQTL** : "methylation Quantitative Trait Locus" pour locus modulant le niveau de méthylation de gènes

Table des matières

Première partie

Contexte & Motivations

1 L'épidémiologie génétique

L'épidémiologie est l'étude, au sein de populations, de la fréquence, de la répartition et des facteurs de risques des maladies. Les études épidémiologiques sont à but descriptif (pour décrire et quantifier un phénomène, par exemple pour mesurer la prévalence d'une maladie), évaluatif (pour déterminer l'intervention ou le traitement le plus efficace) ou étiologique (pour établir les relations entre l'apparition d'une maladie et ses facteurs de risque). Il existe différents types d'études à but étiologique en fonction de la méthodologie employée que ce soit au niveau de la chronologie du recueil des données ou du type de comparaison effectuée : les enquêtes transversales, les enquêtes cas-témoins et les enquêtes de cohorte. L'enquête transversale étudie tous les individus présents au moment de l'enquête avec un recueil rétrospectif des données sur l'exposition. Elle permet d'estimer la prévalence d'exposition et de la maladie ainsi que le risque relatif et l'odds ratio. Ce type d'étude est mal adapté lorsque l'exposition ou la maladie sont rares et sont plus sujettes aux biais. Une enquête cas-témoins est basée quant à elle sur l'étude de deux groupes de sujets : un groupe contenant des sujets atteints par la maladie, l'autre groupe contenant des sujets sains issus de la même population que les cas et servant de témoins. Dans ce type d'enquête, la sélection est réalisée en fonction de la présence ou non de la maladie et le recueil des expositions est toujours rétrospectif. Ce type d'étude est adapté aux maladies rares car elles sont "enrichies" en sujets atteints. Enfin, l'enquête de cohorte est basée sur l'étude d'un groupe de sujets suivis (une cohorte) dans le temps (étude longitudinale) dans le but de mesurer l'incidence d'une maladie et le risque relatif lorsque l'on compare deux cohortes (l'une exposée et l'autre non exposée par exemple). La sélection des sujets se base sur leur niveau d'exposition aux facteurs de risques étudiés. Le recueil des données est le plus souvent prospectif mais peut également être historique. Ce type d'étude est particulièrement adapté lorsque l'exposition est rare et permet d'estimer le taux d'incidence, de risque et de prévalence d'exposition.

L'épidémiologie génétique est la partie de l'épidémiologie qui s'intéresse à l'identification des déterminants génétiques des maladies et des traits quantitatifs associés, comme leurs facteurs de risque. Pendant de nombreuses années, la recherche de nouveaux facteurs de risque génétiques a beaucoup reposé sur l'approche dite "gènes candidats" qui consiste à tester l'association entre une maladie (ou un trait biologique quantitatif) et des polymorphismes génétiques au sein de gènes connus dont les fonctions biologiques font penser qu'ils pourraient être impliqués dans la pathologie étudiée. Cette approche s'est révélée efficace pour découvrir de nouveaux polymorphismes associés aux pathologies

humaines. Le principal inconvénient de cette approche est qu'elle ne pas permet pas de découvrir de nouvelles voies métaboliques impliquées dans les pathologies étudiées. L'identification de nouvelles voies physiopathologiques nécessite d'utiliser des approches agnostiques dites "génome entier".

Les premières approches génome entier furent les analyses de liaison dont l'objectif était de trouver des régions du génome susceptibles de contenir un ou plusieurs variants responsables d'une pathologie, principalement mendélienne, via l'étude de la transmission au sein de familles étendues de variables génétiques (microsatellites, SNP, etc). Ces analyses furent principalement employées lorsque l'on disposait de données familiales sur de grands pedigrees, des jumeaux ou des paires de germains, génotypés pour un nombre relativement restreint de variations génétiques.

Grâce à une meilleure connaissance de la variabilité du génome humain, à l'amélioration des techniques de génotypage et à la baisse de leur coût, la recherche en génétique a subi une révolution avec l'apparition des analyses d'association génome entier (GWAS pour *Genome-Wide Association Study*). En utilisant des puces à ADN permettant de génotyper des centaines de milliers de polymorphismes de simple substitution (*SNP* en anglais), l'approche GWAS consiste à tester l'association entre ces SNPs et un phénotype d'intérêt (une maladie ou un trait quantitatif) généralement dans des échantillons épidémiologiques de grande taille afin de s'assurer une puissance suffisante pour détecter des effets génétiques modestes. L'idée n'est pas nécessairement d'identifier le(s) polymorphisme(s) fonctionnel(s) responsable(s) de la variabilité phénotypique inter-individuelle observée. Mais en se basant sur l'hypothèse que les SNPs génotypés pourraient être en déséquilibre de liaison (LD) avec le(s) variant(s) fonctionnel(s), l'approche GWAS permettrait d'identifier dans un premier temps le gène/locus responsable. Des milliers d'études GWAS ont été réalisées depuis 2005. Au moment de la rédaction de cette thèse, les bases de données GWAS Central (T. Beck *et al.* 2014) et GWAS Catalog (Welter *et al.* 2014) rapportent respectivement les résultats de 1831 et 2175 études de ce type. Les puces à ADN ont d'abord permis d'étudier les associations entre les polymorphismes communs dont la fréquence de l'allèle mineur est supérieure à 5% dans la population générale. Plus récemment, le développement d'outils d'inférence statistique (communément appelée "imputation") et les données obtenues à partir des projets de séquençage haut-débit de grandes populations (c'est-à-dire les projets "1000 Genomes" et "GoNL"), permettent désormais d'étudier des associations avec des polymorphismes dont la fréquence allélique est de l'ordre de 1%.

Des milliers de polymorphismes ont ainsi pu être trouvés associés à des phénotypes biologiques par les approches GWAS. Cependant, dans la majorité des traits étudiés, les polymorphismes identifiés

n'expliquent qu'une part relativement modeste de la variabilité inter-individuelle des phénotypes et de leur héritabilité (pour faire simple, la part de la génétique dans cette variabilité). Par exemple, une centaine de polymorphismes a été trouvé par étude GWAS associée à la variabilité inter-individuelle de l'indice de masse corporelle (Locke *et al.* 2015). Et ces polymorphismes expliquent moins de 3% de cette variabilité. De même, seuls 12% de la variabilité des taux plasmatiques de lipides est expliquée par les 150 polymorphismes identifiés (Global Lipids Genetics Consortium 2013).

Le développement des puces à ADN s'est ensuite accompagné du développement des puces à ARN permettant de mesurer le niveau d'expression de l'ensemble des gènes exprimés dans un tissu ou une cellule (ou un mélange de cellules), donnant lieu aux études d'expression génome entier (GWES pour *Genome-Wide Expression Study*). Les études GWES ont obtenu moins de succès que les études GWAS pour identifier directement de nouveaux gènes de susceptibilités aux maladies complexes lorsque les expressions géniques étaient mises en relation avec des phénotypes d'intérêt. Elles ont en revanche permis d'identifier de nouveaux réseaux de gènes et voies biologiques associées à des maladies (Heinig *et al.* 2010 ; Horvath *et al.* 2006). De plus, la combinaison des données génétiques issues des GWAS et celles d'expression issues de GWES a permis de raffiner certaines associations observées dans les GWAS en identifiant parmi les SNPs identifiés ceux qui pouvaient influencer l'expression de gènes (eSNPs), et donc être des "candidats fonctionnels" (Rotival *et al.* 2011).

Une part importante de la variabilité génétique restant inconnue, la communauté scientifique se tourne à présent vers des nouveaux champs de recherche comme le microbiote (notamment le microbiote intestinal) (T. H. M. P. Consortium 2012 ; Turnbaugh, Hamady *et al.* 2009 ; Turnbaugh, Ley *et al.* 2006), les métabolites (composés organiques issus du métabolisme) (Sabatine *et al.* 2005 ; Sreekumar *et al.* 2009) ou les phénomènes épigénétiques avec notamment l'étude de la méthylation de l'ADN. Cette dernière suscite beaucoup d'intérêt depuis 2010 avec l'apparition, là encore, de puces de haute densité permettant d'étudier la méthylation de l'ADN à grande échelle, donnant lieu aux études d'associations méthylome entier ou *Methylome-Wide Association Studie* (MWAS) en anglais (voir la partie décrivant les études MWAS, p. 91). La terminologie EWAS pour *Epigenome-Wide Association Studie* est parfois également utilisée pour faire référence à ce type d'analyse. Suivant une philosophie très similaire à celle de l'approche GWAS, l'approche MWAS consiste à tester l'association entre les niveaux de méthylation de l'ADN tout au long du génome et un phénotype d'intérêt. Il est également possible de combiner les données issues d'une MWAS avec celles issues des GWAS pour identifier des polymorphismes influençant les niveaux de méthylation (mSNPs) au moyen d'études

mQTL pour analyse *methylation Quantitative Trait Loci*. L'intégration avec des données d'expression est également possible pour identifier des sites de méthylation influençant l'expression de gène (eQTM pour *expression Quantitative Trait Methylation*) (Gutierrez-Arcelus *et al.* 2013). La figure 1 schématise les différentes analyses réalisables en fonction des données disponibles.

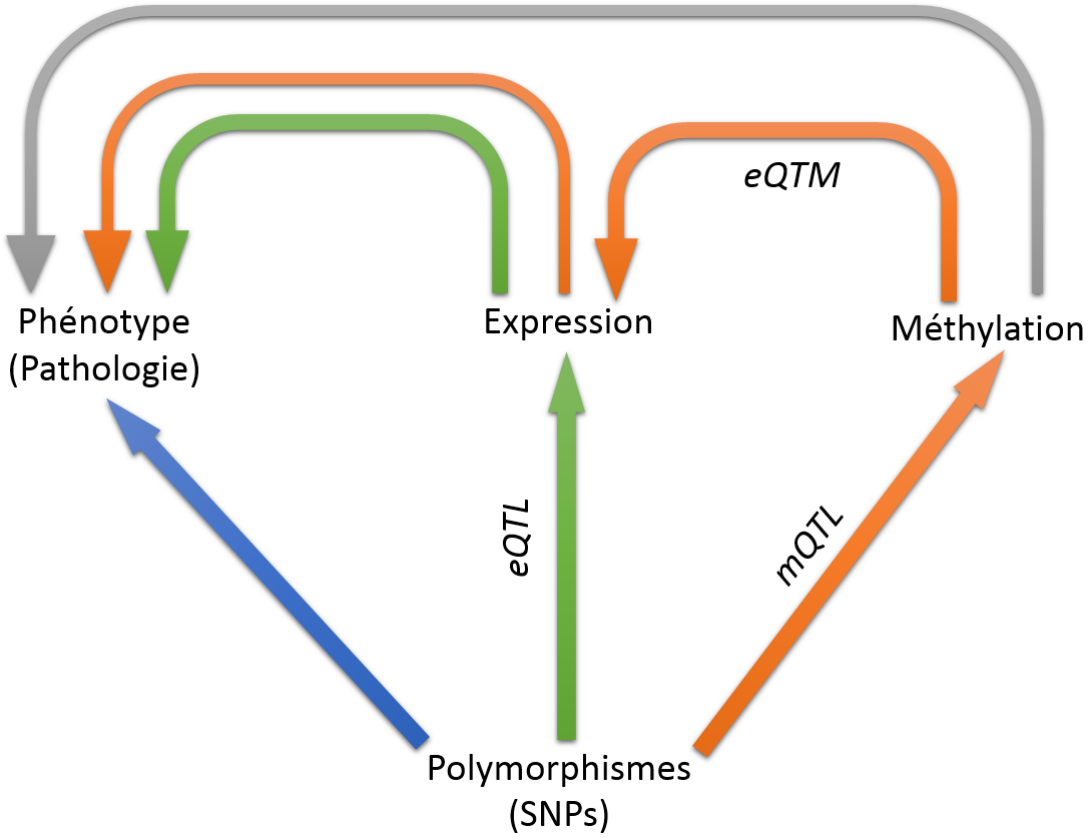

Figure 1 – Représentation schématique de ce qui est possible d'étudier à partir de données GWAS (en bleu), GWES (en vert), MWAS (en orange et gris) en relation avec un phénotype d'intérêt.

2 La méthylation de l'ADN

La méthylation de l'ADN est une modification réversible de l'ADN génomique, plus précisément un mécanisme épigénétique qui consiste en l'ajout d'un groupe méthyle (CH_3 : 1 atome de Carbone lié à 3 atomes d'Hydrogène) au niveau du carbone 5 d'une cytosine (c'est-à-dire au niveau de l'atome de carbone en 5ème position du cycle d'une cytosine), voir la figure 2.

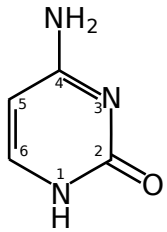

FIGURE 2 – Cytosine avec la nomenclature chimique

Cette modification est réalisée le plus souvent au sein d'une séquence 5'-Cytosine-phosphate-Guanosine, communément appelée dinucléotide CpG ou site CpG. Au-delà de l'ADN, le mécanisme de méthylation peut également toucher certains ARNs (MEYER *et al.* 2012) ou même certaines protéines comme les histones qui peuvent l'être au niveau de leurs acides aminés arginine (BEDFORD & RICHARD 2005) ou lysine (MARTIN & Y. ZHANG 2005). Ce travail de thèse porte uniquement sur la méthylation des cytosines au sein d'un site CpG.

2.1 Les îlots CpG

La répartition des sites CpG au sein du génome n'est pas aléatoire mais répond à une structure particulière (SANDOVAL *et al.* 2011). La majorité des sites CpG se situent dans des îlots CpG (*CpG islands* ou *CGI* en anglais), c'est à dire une succession de sites CpG. Les îlots sont des régions enrichies à approximativement 60 à 70% de sites CpG et possèdent une taille d'environ 850 à 1800 paires de bases (LANDER *et al.* 2001). Il est communément admis que ces régions sont généralement non méthylées dans tous les types cellulaires. Les extrémités de ces îlots, appelées rivages CGI (*CGI shores* en anglais), sont moins enrichies que les îlots et s'étendent sur environ 2 000 paires de bases part et d'autre de l'îlot. Contrairement aux îlots CpG qui sont peu méthylés, les rivages CGI voient une augmentation de la variabilité du niveau de méthylation. Viennent ensuite les bords CGI (*CGI shelves* en anglais), encore moins enrichis que les rivages CGI et qui s'étendent eux de 2 000 à 4 000 paires de bases de l'îlot. Il est également possible de trouver des sites CpG isolés des îlots, on parlera alors de pleine mer CGI (*CGI open sea* en anglais). La figure 3 illustre un îlot CpG avec ses contours.

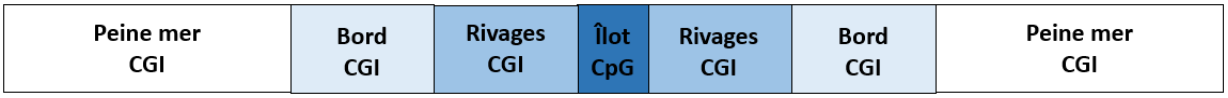

FIGURE 3 – Schéma d'un îlot CpG avec ses contours.

Chez l'homme, environ 60 à 70% des gènes présentent dans leur région promotrice ou leur premier

exon des îlots CpG (Antequera & A. Bird 1993 ; Illingworth & A. P. Bird 2009 ; Miranda & Jones 2007) dont au moins 70% seraient sujets à méthylation (W. Dai *et al.* 2008). Plus précisément, ils chevauchent les régions promotrices de tous les gènes de ménage et d'environ 40% des gènes dont l'expression est restreinte à certains tissus (Illingworth & A. P. Bird 2009 ; Miranda & Jones 2007).

La méthylation de l'ADN est un mécanisme tissu-spécifique, c'est à dire que chaque type cellulaire peut avoir un profil de méthylation différent (Roadmap Epigenomics Consortium *et al.* 2015). Par exemple, les différents types cellulaires composant le sang périphérique possèdent des profils de méthylation différents (H. Ji *et al.* 2010).

3 Les mécanismes de méthylation de l'ADN

La méthylation de l'ADN se fait par le biais des enzymes de la famille des ADN méthyltransférases (DNMT) qui vont catalyser cette réaction d'ajout d'un groupement méthyle. Ces enzymes DNMT vont transférer le groupe méthyle d'une S-adénosylméthionine (SAM) vers la cytosine. Il en résulte une coenzyme appelée S-adénosylhomocystéine (SAH) (figure 4).

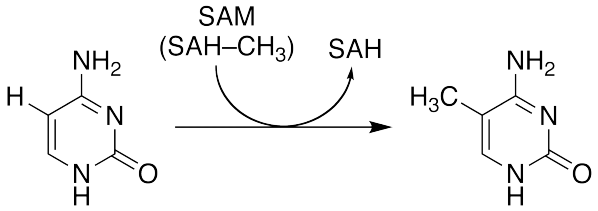

Figure 4 – Schéma de la méthylation

3.1 Méthylation *de novo*

La méthylation de l'ADN peut apparaître sur un site CpG sans qu'il n'y ait d'antécédent de méthylation sur ce site, c'est à dire sur de l'ADN dépourvu de méthylation. Ce mécanisme de méthylation *de novo* fait intervenir les enzymes DNMT3A et DNMT3B (Okano *et al.* 1999) qui, en combinaison avec DNMT3L (Bourc'his *et al.* 2001), vont catalyser l'ajout d'un groupement méthyle sur la cytosine des sites CpG. Les enzymes DNMT sont capables de se fixer à l'ADN via leur domaine PWWP mais la manière dont les enzymes DNMT3A et DNMT3B ciblent les régions à méthyler n'est pas totalement connue bien que certaines hypothèses aient été proposées. La première hypothèse fait appel aux ARNs interférents (ARNi) qui cibleraient les DNMT pour ainsi réprimer certaines séquences

de l'ADN (K. V. Morris *et al.* 2004). L'hypothèse principale fait appel quant à elle aux facteurs de transcription qui se lieraient à des séquences spécifiques de l'ADN et recruteraient les DNMT pour méthyler l'ADN (Brenner *et al.* 2005) ou au contraire pour la protéger de la méthylation (Lienert *et al.* 2011). Les îlots CpG semblent être protégés de la méthylation essentiellement par des facteurs de transcription (Gebhard *et al.* 2010). Or, lorsqu'un site de fixation d'un facteur de transcription est muté, le facteur de transcription est incapable de se fixer et donc de protéger les îlots CpG. Ceux-ci seraient alors sujets à la méthylation (Brandeis *et al.* 1994).

La méthylation *de novo* pourrait donc se faire par la combinaison de deux mécanismes : les enzymes DNMT3A et DNMT3B pourraient premièrement être recrutées par des facteurs de transcription spécifiques pour méthyler ou non les sites CpG et deuxièmement elles pourraient simplement méthyler tous les sites CpG du génome non protégés.

3.2 Méthylation de maintenance

Le génome contient de l'ADN hémiméthylé, autrement dit, un seul des deux brins d'ADN est méthylé. Ce phénomène est particulièrement important lors de la phase S du cycle cellulaire. En effet lors de cette phase, les ADN polymérases vont répliquer l'ADN. Au cours de ce processus, les deux brins de l'ADN vont être séparés, ce qui va permettre l'accès des ADN polymérases qui vont lire un brin et générer un brin complémentaire dépourvu de méthylation. On sera donc en présence d'ADN hémiméthylé. Les enzymes DNMT1 ciblent les régions du génome hémiméthylées (Yokochi & Robertson 2002) et catalysent la réaction d'ajout du groupement méthyle sur le brin nouvellement synthétisé. Ce mécanisme explique notamment pourquoi la méthylation de l'ADN est une marque épigénétique transmissible à la descendance et intervient dans l'empreinte parentale.

3.3 Perte de la méthylation

La perte de la méthylation peut se faire de manière passive ou active. La déméthylation passive se déroule lors de la division cellulaire, l'inhibition ou une dysfonction de l'enzyme DNMT1 dont l'un des rôles est de méthyler le brin nouvellement synthétisé lors de la phase S peut aboutir à un brin d'ADN peu ou pas méthylé. En ce qui concerne la déméthylation active, elle peut se dérouler à n'importe quelle phase du cycle cellulaire et requiert l'intervention de réactions enzymatiques, des successions d'oxydations et de désaminations. Il n'existe pas à ce jour dans les cellules de mammifère de mécanismes connus pour cliver la forte liaison carbone-carbone qui lit la méthylation à la cytosine.

L'astuce mise en place est de transformer la cytosine méthylée en un élément reconnu par les systèmes de correction de l'ADN (réparation par excision de base, *BER* en anglais) qui vont remplacer la cytosine modifiée par une cytosine non méthylée. La modification des cytosines méthylées peut se faire sur deux sites : soit le groupement méthyle, soit le groupement amine.

Les cytosines méthylées sont reconnues par les protéines de la famille TET (*ten-eleven translocation*) qui va les oxyder, ce qui va les transformer en cytosines hydroxyméthylées. Les cytosines hydroxyméthylées peuvent subir de nouvelles oxydations pour former des cytosines formylées ou des cytosines carboxylées en fonction du nombre d'oxydations subit (respectivement une ou deux oxydations successives). Une fois transformées en cytosines formylées ou carboxylées, l'enzyme *thymine-DNA glycosylase* (TDG) va les reconnaître et les remplacer par des cytosines non méthylées. Une désamination des cytosines méthylées et hydroxyméthylées est également possible via le complexe enzymatique AID/APOBEC (pour *activation-induced cytidine deaminase/apolipo-protein B mRNA-editing enzyme complex*). Il en résultera respectivement des thymines et des uraciles hydroxyméthylées qui pourront également être remplacées via l'enzyme *thymine-DNA glycosylase* par des cytosines non méthylées. Les cytosines hydroxyméthylées, formylées, et carboxylées peuvent également subir comme les cytosines méthylées une perte passive de leur marque épigénétique lors de la division cellulaire.

Le phénomène de désamination explique la raison de la sous-représentation de site CpG dans le génome, environ 20% de ce que l'on pourrait attendre aléatoirement (Lander *et al.* 2001). En effet, les cytosines non méthylées peuvent subir une désamination qui va les convertir en uracile, base azotée propre à l'ARN. Celle-ci sera détectée par l'enzyme *uracile-DNA glycosylase* qui la corrigera. Or, lorsque la cytosine qui subit une désamination est méthylée, celle-ci sera convertie en thymine (A. P. Bird 1980), base azotée propre à l'ADN et donc moins facilement détectable par les mécanismes de correction de l'ADN comme les enzymes *thymine-DNA glycosylase* et *methyl-CpG-binding domain protein 4* (Hendrich, Hardeland *et al.* 1999). La figure 5 représente les différents mécanismes de méthylation ou de déméthylation des cytosines.

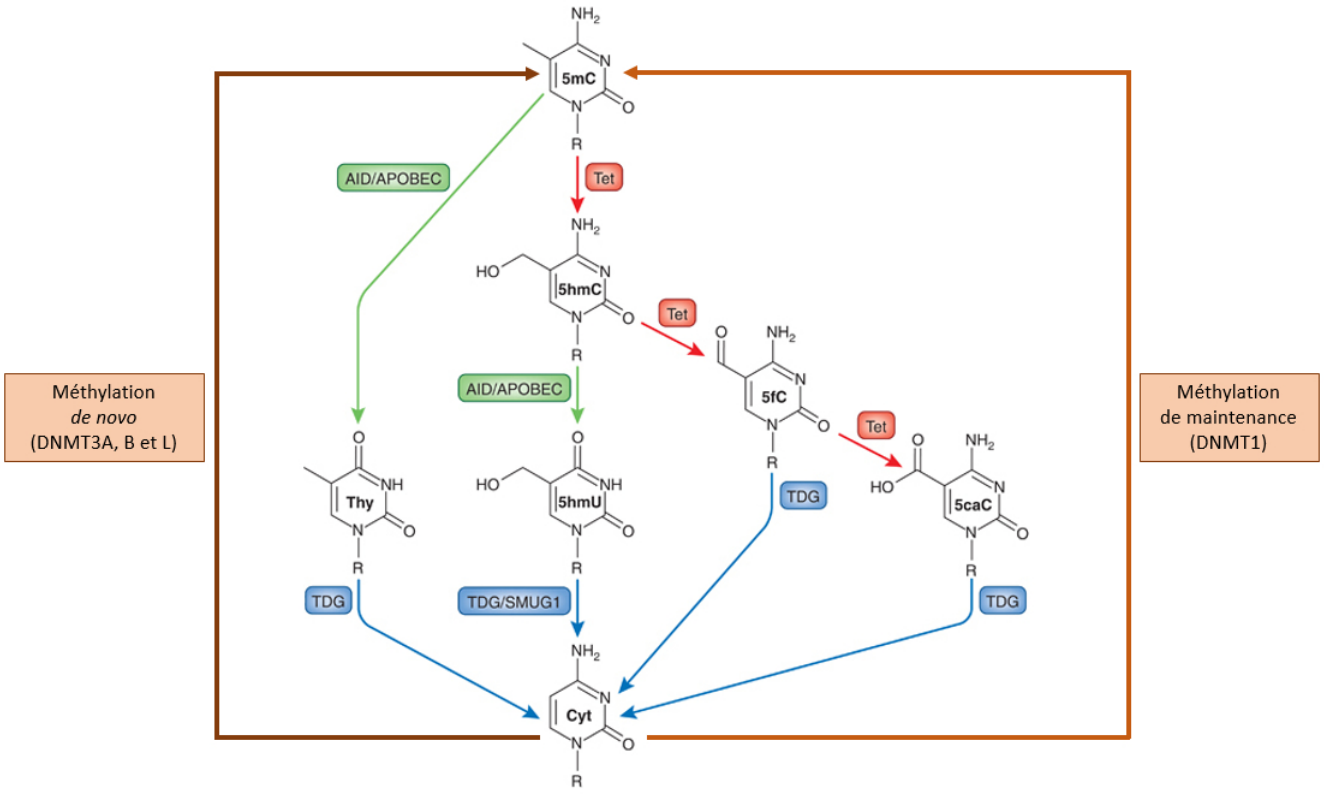

FIGURE 5 – Mécanismes de la méthylation (marron) et de déméthylation active dont les oxydations (rouge), les désaminations (vert) et les remplacements par l'enzyme TDG (bleu). Schéma basé sur les figures de L. D. MOORE *et al.* 2013 et de SCHÜBELER 2015.

4 Les fonctions de la méthylation de l'ADN

L'une des principales fonctions de la méthylation est de pouvoir réguler l'expression des gènes. Les mécanismes par lesquels la méthylation régule l'expression génique sont encore mal compris. Le caractère non-méthylé d'une cytosine est généralement considéré comme l'état "naturel" de base du gène associé lui permettant d'être prêt pour la transcription. À l'inverse, la méthylation de l'ADN d'une région régulatrice va rendre cette dernière plus ou moins accessible aux facteurs de transcription et va attirer des enzymes qui vont modifier localement la structure et la composition de la chromatine, ces deux mécanismes pouvant induire une dis-régulation de l'expression du gène associé. Récemment, plusieurs études comparant le niveau de méthylation de l'ADN et l'expression des gènes démontrent que ce mécanisme est plus complexe et qu'une hyperméthylation peut également être associée à une augmentation de l'expression du gène (EIJK *et al.* 2012 ; GUTIERREZ-ARCELUS *et al.* 2013). Les auteurs de ces études suggèrent que la position du site CpG dans le gène est importante, ainsi la méthylation de l'ADN dans une région promotrice induit une inhibition de l'expression génique alors

que la méthylation de l'ADN dans le corps du gène (Jjingo *et al.* 2012) ou dans la région 3' UTR (Grundberg *et al.* 2013) induirait une augmentation de l'expression génique.

La méthylation de l'ADN est impliquée dans de nombreux autres processus biologiques (Robertson 2002) comme le développement, la différenciation cellulaire, la régulation de la transcription des gènes, la modulation de la structure de la chromatine, le maintien de la stabilité du génome, l'inhibition de séquences répétées ou d'éléments d'ADN parasitaires (comme les rétrovirus endogènes) (Walsh *et al.* 1998), l'inactivation du chromosome X ou plus généralement l'empreinte parentale (Li *et al.* 1993). Elle intervient dans ces divers processus biologiques via deux familles de protéines, la famille des protéines à domaine MBD (Q. Du *et al.* 2015) qui reconnaissent les sites CpG méthylés et la famille des protéines à domaine ZF-CxxC (H. K. Long *et al.* 2013) qui, à l'inverse, reconnaissent les sites CpG non méthylés.

4.1 Les protéines à domaine MBD

La famille des protéines MBD pour *Methyl-CpG binding domain* est actuellement composée d'une dizaine de protéines contenant toutes un domaine MBD au sein de leur séquence (Q. Du *et al.* 2015 ; Hendrich & Tweedie 2003). Le domaine MBD est composé de 70 à 85 acides aminés et permet aux protéines le possédant de reconnaitre les sites CpG méthylés. La première protéine connue pour avoir la capacité de se fixer sur un site CpG méthylé via son domaine MBD est la protéine MeCP2. Lorsque MeCP2 se fixe sur un site CpG méthylé, elle va recruter des partenaires pouvant modifier les marques épigénétiques des histones environnantes comme supprimer les marques d'acétylation ou ajouter des marques de méthylation. La modification de ces marques épigénétiques des histones permet d'inhiber la transcription via des remodelages du nucléosome permettant de maintenir l'état condensé de la chromatine. En effet, l'hyperacétylation [1] et l'hypométhylation des histones sont associées à l'euchromatine, c'est à dire une chromatine décondensée et prête pour la transcription tandis qu'à l'inverse, l'hypoacétylation et l'hyperméthylation des histones sont associées à l'hétérochromatine, c'est à dire une chromatine condensée ne permettant pas à l'ADN d'être transcrit. La protéine MBD1 contient à la fois un domaine MBD lui permettant de se fixer sur des sites CpG méthylés mais également trois domaines ZF-CxxC lui permettant de se fixer sur des sites CpG non méthylés. Tout comme MeCP2, MBD1 peut recruter des partenaires pouvant enlever l'acétylation ou ajouter la méthylation aux histones pour condenser la chromatine et ainsi inhiber la transcription. Les autres

1. L'acétylation est l'ajout d'un groupement acétyle (COCH3) sur une protéine, ici sur une histone.

protéines de la famille MBD ont les mêmes principes d'actions avec des différences au niveau de leur domaine MBD et des partenaires recrutés ce qui leur permet d'avoir des caractéristiques des fixations et fonctions différentes tout en gardant globalement le rôle d'inhiber la transcription (Q. Du *et al.* 2015).

Si les domaines MBD (pour *Methyl-CpG binding domain*) permettent aux protéines de se fixer sur les sites CpG méthylés, les domaines ZF-CxxC (*Unmethylated-CpG binding zinc finger*) permettent quant à eux de se fixer sur les sites CpG non méthylés.

4.2 Les protéines à domaine ZF-CxxC

La famille des protéines à domaine ZF-CxxC (domaine en doigts de zinc de type CxxC) est actuellement composée d'une dizaine de protéines qui, via leur domaine ZF-CxxC, peuvent se fixer sur les sites CpG non méthylés pour potentiellement renforcer l'état non méthylé ou recruter d'autres protéines pour déméthyler des îlots CpG. Il existe trois sous-groupes de domaine ZF-CxxC composés de 35 à 42 acides aminés et contenant deux clusters riches en cystéines et conservés au cours de l'évolution. Les protéines contenant le premier sous-groupe du domaine ZF-CxxC, dont notamment KDM2A et KDM2B, se fixeraient sur les îlots CpG non méthylés. La protéine KDM2A qui contient un domaine ZF-CxxC, contient également un domaine JmjC (Jumonji C) qui catalyse la suppression de la méthylation des lysines de l'histone H3, en particulier la diméthylation de la lysine 36 de l'histone H3 (H3K36me2) dont le rôle est d'inhiber la transcription. Il est donc tout naturel de faire l'hypothèse que, via son domaine ZF-CxxC, KDM2A va donc déméthyler les H3K36me2 au sein d'un îlot CpG non méthylé, ce qui va entraîner l'apparition d'un environnement de la chromatine favorable à l'initialisation de la transcription. La protéine KDM2B possède le même fonctionnement que KDM2A mais aurait, en plus, une préférence pour un sous-groupe d'îlots CpG situés sur des gènes impliqués dans le développement embryonnaire, la morphogenèse et la différenciation cellulaire. Comme ces gènes sont généralement liés à des complexes PRC (pour *polycomb group repressive complexe*) dont le rôle est de méthyler certaines histones pour inhiber la transcription, il est logique de faire l'hypothèse que KDM2B serait impliqué dans l'inhibition induite par les complexes PRC notamment pour les recruter sur les îlots CpG non méthylés. Le second sous-groupe du domaine ZF-CxxC est porté par la protéine MBD1 et ne possède pas de boucle de liaison à l'ADN fonctionnelle. Sa fonction serait plus vraisemblablement de moduler des interactions protéine-protéine. Le dernier sous-groupe du domaine ZF-CxxC quant à lui est notamment présent sur les protéines TET1 et TET3. Il diffère du premier

sous-groupe par une région de liaison à l'ADN différente. Bien que les mécanismes de fixation de ce sous-groupe ne soient pas totalement compris, il semblerait que la différence dans la région de fixation permettrait à la protéine TET1 de se fixer sur les sites CpG quel que soit l'état de méthylation de la cytosine et à la protéine TET3 de se fixer sur des cytosines non méthylées hors d'un site CpG. Pour résumer, le domaine ZF-CxxC des protéines permet à ces dernières de se fixer sur des sites CpG dont la cytosine est non méthylée pour y recruter d'autres protéines pouvant modifier l'état de la chromatine via, par exemple, l'ajout de méthylation sur des histones (H. K. Long *et al.* 2013).

De par les multiples fonctions des protéines et de leurs partenaires qui interagissent au niveau des sites CpG en fonction de leur état de méthylation, des altérations dans la régulation des activités du génome liées à la méthylation de l'ADN peuvent générer des désordres biologiques.

5 L'implication de la méthylation dans des mécanismes physiopathologiques

Les différents troubles biologiques associés des profils aberrants de méthylation de l'ADN couvrent un large spectre de pathologies allant des pathologies développementales aux cancers en passant par diverses pathologies communes multifactorielles (Tost 2009).

Les différentes pathologies développementales (Bjornsson *et al.* 2004 ; Feinberg 2007 ; Robertson 2005) associées à des profils de méthylation particuliers ont des origines différentes. Pour certaines, il s'agit d'un phénomène d'empreinte parentale, pour d'autres une expansion instable de répétitions (Tost 2009). Par exemple, pour des pathologies ayant pour origine un désordre de l'empreinte génétique (l'exemple le plus connu est l'empreinte parentale) comme le syndrome de Beckwith-Wiedemann, le syndrome de Prader-Willi, ou le syndrome d'Angelman (Robertson 2005), le phénomène s'explique par une dérégulation de la modification épigénétique d'un des deux chromosomes ayant normalement pour but de contrôler l'expression différente des gènes d'origine paternelle et maternelle. Un profil aberrant de méthylation de l'ADN peut entraîner un trouble de cette empreinte qui va aboutir à des expressions géniques pathologiques et ainsi aux différents syndromes. Le profil aberrant de méthylation va ainsi induire l'expression de l'allèle normalement éteint ou la répression de l'allèle normalement actif. Pour le syndrome de l'X fragile, qui illustre les maladies à expansions instables de répétitions ayant un profil de méthylation particulier, le gène *FMR1* du chromosome X contient un triplet "CGG" répété moins de 20 fois chez les individus sains. Chez les individus atteints

du syndrome de l'X fragile, ce triplet est répété plus de 200 fois (Oberle *et al.* 1991). Cette région riche en sites CpG sera hyperméthylée chez les personnes atteintes ce qui entraînera une inhibition de la transcription du gène *FMR1*.

La méthylation est également impliquée dans le développement de certains cancers (Esteller 2008 ; Feinberg & Tycko 2004 ; Jones & Baylin 2007) et de pathologies multifactorielles. Ces maladies sont connues pour être sous l'influence de facteurs génétiques mais également environnementaux (Herceg 2007 ; Herceg & Vaissière 2011). Certains de ces facteurs environnementaux peuvent perturber les mécanismes de méthylation de l'ADN. C'est le cas par exemple de la consommation de tabac (Breitling *et al.* 2011 ; Christensen *et al.* 2009 ; Joubert *et al.* 2012 ; Tsaprouni *et al.* 2014 ; Zeilinger *et al.* 2013) et d'alcool (Kaminen-Ahola *et al.* 2010 ; Philibert *et al.* 2012 ; R. Zhang *et al.* 2013) ou de l'exposition à certains polluants comme l'arsenic (Kile *et al.* 2014 ; Koestler, Avissar-Whiting *et al.* 2013).

5.1 Les cancers

Comme nous l'avons vu dans la partie précédente, la méthylation de l'ADN permet de réguler l'expression des gènes. Or, une régulation défectueuse peut aboutir à des effets indésirables et potentiellement pathogènes comme cela peut être le cas avec des gènes régulateurs de la prolifération cellulaire.

Les premières indications d'un lien entre certains cancers et des anomalies de méthylation datent de plus de 30 ans avec la découverte d'une hypométhylation observée dans des tumeurs (Riggs & Jones 1983). Ce résultat a ensuite été retrouvé dans de nombreux types de cancers (Feinberg & Tycko 2004 ; Herceg & Hainaut 2007 ; Jones & Baylin 2007). L'hyperméthylation d'îlots CpG peut induire l'inactivation de gènes suppresseurs de tumeurs (régulateur négatif de la prolifération cellulaire) ce qui entraîne une prolifération anormale de cellules et peut aboutir à l'apparition de tumeurs. C'est le cas par exemple du gène *HIC1* dont l'hyperméthylation est associée à différents types de cancer dont des carcinomes, sarcomes et lymphomes (Burton 2003 ; W. Y. Chen *et al.* 2003 ; Greenwood 2003). Il est également possible qu'une méthylation anormale entraîne l'activation de proto-oncogènes (régulateur positif de la prolifération cellulaire), ce qui aura le même effet d'entraîner une prolifération anormale de cellules. Par exemple, l'hypométhylation de l'oncogène *ELMO3* et l'augmentation de son expression sont associés au développement du cancer du poumon et à la formation de métastases (Søes *et al.* 2014). Un autre exemple plus complexe est l'hypométhylation

du promoteur du gène *LINE-1* qui va induire l'expression d'un transcrit alternatif du proto-oncogène *MET* dans le cancer de la vessie (Wolff *et al.* 2010). Une étude cas-témoins a mis en évidence une hyperméthylation du gène *RASSF1A* (gène suppresseur de tumeurs) dans le cancer du poumon (Vineis *et al.* 2011). Cette même étude a montré que les taux plasmatiques de vitamine B9 sont associés à une augmentation des niveaux de méthylation des gènes *RASSF1A* et *MTHFR*, tandis que les taux sanguins de méthionine sont associés à une diminution de la méthylation du gène *RASSF1A* (Vineis *et al.* 2011).

Certains types de cancers peuvent également avoir des profils de méthylation particuliers. Une étude cas-témoins a montré qu'une hypométhylation globale de l'ADN leucocytaire est associée au cancer de la vessie (L. E. Moore *et al.* 2008). Cette association est également retrouvée dans le cancer des voies aérodigestives supérieures (Hsiung *et al.* 2007). Les mécanismes de cette association sont mal compris mais l'hypothèse proposée est que les cellules ayant un faible taux de méthylation seraient moins stables génétiquement et pourraient plus facilement entrer en apoptose. Les cellules génétiquement modifiées seraient ainsi désavantagées au niveau de leur survie, ce qui serait bénéfique pour l'hôte lorsqu'il est exposé à des agents endommageant l'ADN. Un autre exemple est le cancer du sein qui serait quant à lui associé à une hyperméthylation de 264 loci d'îlots CpG (Hill *et al.* 2011) dans un groupe de gènes comprenant le gène *RECK*, déjà connu pour être associé à d'autres cancers comme celui des voies aérodigestives supérieures (N. K. Long *et al.* 2008) ou des poumons (Chang *et al.* 2007). Ou encore, le carcinome du pénis qui est lui trouvé associé à une hyperméthylation de 171 sites CpG (Kuasne *et al.* 2015).

De plus, il est possible de trouver de nouveaux biomarqueurs à partir de la méthylation de l'ADN provenant de cellules sanguines. En effet, la présence d'un cancer de l'ovaire est associée à un profil de méthylation particulier dans les cellules sanguines (Teschendorff, Menon *et al.* 2009). L'étude de la méthylation dans le contexte de l'oncologie permet également de pouvoir classifier les tumeurs pour pouvoir espérer mieux les traiter par la suite. Dans le cas des cancers colorectaux, il est possible de les différencier en divers sous-groupes en fonction de leur profil de méthylation (Hinoue *et al.* 2012).

Au-delà des cancers, la méthylation peut également intervenir dans des mécanismes physiopathologiques associés à d'autres maladies multifactorielles.

5.2 Les autres maladies multifactorielles

De nombreux travaux récents suggèrent que les mécanismes de méthylation seraient impliqués dans l'apparition et le développement de certaines maladies multifactorielles (Bjornsson *et al.* 2004 ; Feinberg 2007), qu'elles soient d'origines neurologiques, auto-immunes ou même d'origines cardio-vasculaires.

Parmi les maladies neurologiques où des profils particuliers de méthylation de l'ADN ont été identifiés, nous pouvons citer la sclérose en plaque, la dépression, l'épilepsie, potentiellement la narcolepsie et en restant dans le domaine du sommeil, le syndrome d´apnées obstructives du sommeil. Dans le cas de la sclérose en plaque, les gènes *BCL2L2* et *NDRG1*, deux gènes régulant la survie des oligodendrocytes, sont trouvés hyperméthylés et exprimés à un plus faible niveau dans les tissus cérébraux des individus atteints de sclérose en plaque par rapport aux individus de contrôles, tandis que les gènes *LGMN* et *CTSZ*, impliqués dans le processus de protéolyse, sont trouvés hypométhylés et exprimés à un plus fort niveau (Huynh *et al.* 2014). La dépression, quant à elle, a été trouvée associée aux niveaux de méthylation (mesurés à partir des cellules du sang périphérique) de plusieurs sites CpG de gènes précédemment associés à des phénotypes neuropsychiatriques comme le gène *WDR26* trouvé associé à la dépression majeure dans une étude GWAS (Córdova-Palomera *et al.* 2015), mais également de gènes impliqués dans la voie de signalisation des glucocorticoïdes (Córdova-Palomera *et al.* 2015 ; Januar *et al.* 2015). Des niveaux de méthylations différents seraient également retrouvés dans l'hippocampe de patients atteints d'épilepsie du lobe temporal en comparaison à des individus sains (Miller-Delaney *et al.* 2014). Les mécanismes de méthylation pourraient également être impliqués dans la narcolepsie puisque des mutations du locus *DNMT1*, gène codant une enzyme impliquée dans le maintien de la méthylation, seraient associées à la maladie dans certaines familles (Winkelmann *et al.* 2012). En ce qui concerne le syndrome d´apnées obstructives du sommeil, il serait associé à l'augmentation de taux de méthylation des gènes *FOXP3* et *IRF1*, tous deux impliqués dans la réponse immunitaire (Kim *et al.* 2012).

La méthylation est également impliquée dans des maladies auto-immunes (Richardson 2007) comme la polyarthrite rhumatoïde où le niveau de méthylation de gènes du complexe majeur d'histocompatibilité (système de reconnaissance du soi) serait différent entre les cas et les témoins (Y. Liu *et al.* 2013). Le taux plasmatique d'immunoglobuline E, protéine de la réponse immunitaire impliquée dans les maladies allergiques comme l'asthme ou l'eczéma, est associé à des profils de méthylation

particuliers (Liang *et al.* 2015). Il en est de même pour le diabète de type 1, où la méthylation de 132 sites CpG est associée à l'apparition de la maladie chez 15 paires de jumeaux (Rakyan, Beyan *et al.* 2011). Une différence de méthylation sur 19 sites CpG a également été trouvée entre des patients atteints de diabète de type 1 sans signe de complication rénale et des patients atteints de diabète de type 1 ainsi que de néphropathie diabétique (C. G. Bell, Teschendorff *et al.* 2010).

Les maladies cardiométaboliques ne sont pas épargnées, comme le diabète de type 2 où l'haplotype de susceptibilité à l'obésité du locus *FTO* est associé à une augmentation de la méthylation sur ce même locus (C. G. Bell, Finer *et al.* 2010). Des différences de méthylation, pour trois gènes en relation avec l'angiogenèse, sont également retrouvées dans les ventricules gauches de patients atteints de cardiomyopathie idiopathique comparativement à des individus sains (Movassagh *et al.* 2010). Et enfin, la méthylation est également connue pour intervenir dans des mécanismes liés aux pathologies cardiovasculaires comme l'athérosclérose, l'inflammation (Wierda *et al.* 2010) ou d'autres mécanismes (Baccarelli *et al.* 2010 ; Matouk & Marsden 2008). De plus, il a été montré que des facteurs sociologiques comme le statut socio-économique sont également associés à des profils particuliers de méthylation de certains gènes dont des gènes régulateurs de l'inflammation (Stringhini *et al.* 2015). Ce phénomène pourrait expliquer l'apparition plus fréquente de maladies multifactorielles chroniques dans certaines classes sociales.

5.3 L'intérêt d'étudier la méthylation des cellules circulantes

Bien que les marques de méthylation soient spécifiques aux types cellulaires, il n'est pas toujours facile d'avoir accès à l'échelle épidémiologique aux tissus biologiques pertinents pour une maladie d'intérêt. L'accès à des tissus comme le foie ou une tumeur nécessite un acte invasif comme une biopsie. Il est bien plus facile à l'échelle épidémiologique d'avoir accès à des échantillons sanguins qui nécessitent une simple prise de sang. Les échantillons sanguins sont composés des cellules circulantes, c'est-à-dire différents types cellulaires dont principalement les cellules sanguines comme les érythrocytes (également connus sous le nom de globules rouges), les leucocytes (globules blancs) et les thrombocytes (plaquettes sanguines). Les érythrocytes et les thrombocytes ne possédant pas de noyau, l'ADN extrait des cellules circulantes provient donc essentiellement des leucocytes. Bien que les différents types de leucocytes aient des profils de méthylation différents (Y. V. Sun *et al.* 2010), de nombreux travaux ont montré que les niveaux de méthylation observés au sein de l'ADN de sang périphérique pouvaient servir à identifier de nouveaux biomarqueurs de risque de maladie (C. G.

Bell, Finer *et al.* 2010 ; Rakyan, Beyan *et al.* 2011 ; Wierda *et al.* 2010). Ce qui permettrait de pouvoir améliorer le dépistage ou le diagnostic de ces dernières. L'intérêt des cellules circulantes dans le contexte des pathologies cardiovasculaires est d'autant plus grand que les leucocytes sont composés en partie par les lymphocytes, cellules du système immunitaire impliquées dans l'inflammation et donc potentiellement pertinentes pour les pathologies cardiovasculaires (Bhat *et al.* 2013). De plus, lorsque l'ADN est stocké dans des conditions idéales, comme c'est le cas dans de nombreuses biobanques constituées dans le cadre de larges études épidémiologiques, la marque de méthylation de l'ADN est stable dans le temps (Talens *et al.* 2010).

Au-delà des pathologies cardiovasculaires, il a déjà été montré qu'il était possible de trouver de nouveaux biomarqueurs à partir de la méthylation de l'ADN des cellules sanguines dans certains types de cancers (Teschendorff, Menon *et al.* 2009). L'hypothèse à l'origine de ce travail de thèse est que l'étude du méthylome circulant pourrait permettre d'identifier de nouveaux mécanismes épigénétiques impliqués dans la survenue des pathologies cardiovasculaires et permettre de trouver de nouveaux biomarqueurs et de nouvelles cibles thérapeutiques. En effet, actuellement deux familles de molécules ont été développées pour interagir avec l'épigénétique, une famille qui inhibe les déacétylases d'histone (HDACi) et une famille plus intéressante dans le cas de la méthylation de site CpG, qui inhibe les ADN méthyltransférases (DNMTi). L'identification de profils de méthylation aberrants responsables de certaines pathologies pourrait donc permettre une meilleur prise en charge thérapeutique via ces molécules qui agissent sur l'épigénétique dans le but de supprimer des marquages anormaux (Esteller 2007 ; Rasool *et al.* 2015).

6 Les méthodes de mesure de la méthylation de l'ADN

Il est possible de mesurer le niveau de méthylation du génome via différentes méthodes (S. Beck 2010 ; Laird 2010 ; Rakyan, Down *et al.* 2011 ; Tost 2009). Ces méthodes de détection du statut de méthylation se différencient en fonction de technologies et propriétés de biologie moléculaire employées (ou prétraitement). Certaines utilisent des mécanismes de digestion enzymatique sensibles à l'état de méthylation, d'autres les propriétés de fixation de certaines protéines ou anticorps ou encore la conversion au bisulfite de sodium.

Les méthodes se servant des mécanismes de digestion enzymatique utilisent des enzymes de restrictions sensibles à l'état de méthylation qui coupent la molécule d'ADN sur un site spécifique. Parmi les

enzymes les plus utilisées, nous pouvons citer les couples HpaII/MspI et SmaI/XmaI comme exemples. L'enzyme HpaII et son isoschizomère [2] MspI reconnaissent la séquence "CCGG" (Waalwijk & Flavell 1978). HpaII coupe la séquence "CCGG" lorsque la cytosine du site de coupure est non méthylée alors que MspI coupe la séquence "CCGG" quel que soit son état de méthylation. En ce qui concerne l'enzyme SmaI et son néoschizomère [3] XmaI, elles reconnaissent quant à elles la séquence "CCCGGG" (Withers & Dunbar 1993). SmaI coupe la séquence "CCCGGG" lorsque la cytosine du site de coupure est non méthylée tandis que XmaI coupe elle la séquence "CCCGGG" quel que soit son état de méthylation mais est moins efficace lorsque la cytosine du site de coupure est méthylée. Pour obtenir le statut de méthylation, il suffit ensuite de comparer les fragments obtenus par la digestion des deux enzymes d'un même couple. Les méthodes se basant sur l'enrichissement par affinité utilisent soit des anticorps monoclonaux spécifiques aux cytosines méthylées, soit des protéines ayant des domaines qui interagissent avec les sites CpG en fonction de leur état de méthylation comme les protéines MECP2 et MBD2. L'immunoprécipitation induite par les anticorps ou la fixation par les protéines interagissant avec les sites CpG vont sélectionner les fragments d'ADN méthylés qui seront par la suite amplifiés. On obtiendra donc un enrichissement de fragments d'ADN méthylés. Enfin, le traitement au bisulfite de sodium va convertir les cytosines non méthylées en uraciles [4] et laisser intacte les cytosines méthylées. La comparaison des séquences avant et après le traitement au bisulfite de sodium permet d'obtenir le statut de méthylation.

Ces différentes méthodes de détection du statut de méthylation sont employées via différentes technologies comme le séquençage haut débit (*NGS* pour *Next-Generation Sequencing*) ou les puces à ADN. Voici quelques exemples de combinaison méthode/technologie couramment utilisées :

- MeDIP-seq : méthode de séquençage basée sur l'immunoprécipitation (Weber *et al.* 2005) ;
- MethylCap-seq : méthode de séquençage basée sur la capture par affinité avec le domaine MBD de la protéine MeCP2 (Brinkman *et al.* 2010) ;
- MBD-seq : méthode de séquençage également basée sur la capture par affinité avec le domaine MBD de la protéine MBD2 (Serre *et al.* 2010) ;
- WGBS (pour *Whole-Genome Bisulfite Sequencing*) : méthode de séquençage basée sur la conversion au bisulfite de sodium (Lister *et al.* 2009) ;
- RRBS (pour *Reduced Representation Bisulfite Sequencing*) : méthode de séquençage basée sur

2. Enzymes de restriction reconnaissants la même séquence cible et la coupants de façon identique.
3. Enzymes de restriction reconnaissants la même séquence cible et la coupants de façon différente.
4. Une uracile est une base azotée propre à l'ARN qui s'apparie avec la thymine tout comme l'adénine dans l'ADN.

la digestion enzymatique et la conversion au bisulfite de sodium (Gu *et al.* 2011) ;

— Infinium HumanMethylation : méthode utilisant les puces à ADN et basée sur la conversion au bisulfite de sodium (Bibikova, Barnes *et al.* 2011 ; Sandoval *et al.* 2011) ;

— CHARM : méthode utilisant les puces à ADN et basée sur la digestion enzymatique (Irizarry, Ladd-Acosta *et al.* 2008).

Chaque méthode a ses avantages et ses inconvénients (S. Beck 2010 ; Laird 2010 ; Rakyan, Down *et al.* 2011). Certaines couvrent des régions plus importantes du génome que d'autres (séquençage *versus* puces à ADN). Certaines sont plus précises et ont une meilleure reproductibilité (conversion au bisulfite de sodium *versus* enrichissement par affinité), etc. Actuellement la méthode de référence (ou *Gold standard*) à l'échelle du génome est le WGBS (S. Beck 2010), qui remplace une méthode de pyroséquençage (Tost & Gut 2007). Cette nouvelle méthode de référence nécessite une plus grande quantité d'ADN ainsi que des analyses bioinformatiques plus complexes et donc possède un coût plus élevé. Au moment où j'ai débuté ma thèse, la société Illumina venait de développer une puce à ADN permettant de mesurer les niveaux de méthylation d'environ 450 000 sites CpG tout au long du génome, la puce HumanMethylation450k. Cette technologie que je vais décrire en détail dans la partie suivante semblait être un bon compromis pour rechercher des marques de méthylation à partir d'échantillons sanguins collectés au sein d'études épidémiologiques. L'objectif du projet de thèse qui m'a été confié était de déterminer si l'application de cette puce à un échantillon d'ADN sanguin de 350 sujets de l'étude MARTHA (voir p. 22) pouvait permettre de mettre en évidence de nouveaux mécanismes épigénétiques associés à des biomarqueurs du risque cardiovasculaire. Bien que ce type d'étude soit relativement courant de nos jours, ce thème était novateur au début de mon travail de thèse. Peu d'études du méthylome provenant des cellules du sang périphérique étaient réalisées à échelle épidémiologique.

Deuxième partie

Mesure du méthylome par biopuce

Chapitre 1

Les données épidémiologiques utilisées

Les données disponibles pour ce projet de thèse ont été générées à partir de deux études : 350 sujets proviennent de l'étude MARTHA (Oudot-Mellakh *et al.* 2012) et 227 sujets proviennent de l'étude F5L-Pedigrees (Antoni *et al.* 2010). Parmi les 577 individus composants ces deux études, 11 ont été mesurés deux fois sur des puces et lots différents pour évaluer la reproductibilité des mesures et leur concordance. Les 588 échantillons des deux études ont été analysés simultanément et par la même personne au centre de génomique appliquée (*TCAG* pour *The Center for Applied Genomics*) de Toronto au Canada en utilisant le protocole ci-dessous.

L'ADN génomique a été isolé à partir des cellules provenant du sang périphérique des individus en utilisant une adaptation de la méthode proposée par Miller *et al.* 1988. Pour chaque échantillon, 1 µg d'ADN a été converti en utilisant le kit Qiagen EpiTect 96 Bisulfite. Puis, 200 ng d'ADN convertis au bisulfite à 50 ng/µl ont été indépendamment amplifiés, marqués et hybridés sur la puce HM450k. Et enfin les puces ont été scannées par l'appareil Illumina iScan configuré avec les paramètres par défaut.

1 L'étude MARTHA

L'étude MARTHA (MARseille THrombosis Association) (Oudot-Mellakh *et al.* 2012 ; Trégouët *et al.* 2009) a été mise en place par le Pr. Pierre-Emmanuel Morange en 1994 à Marseille et financée par un Programme Hospitalier de Recherche Clinique. L'objectif de cette étude épidémiologique est de découvrir de nouveaux facteurs de risque génétiques de la maladie thrombo-embolique veineuse (MTEV). L'étude est constituée de patients recrutés consécutivement au Laboratoire d'hé-

matologie de l'Hôpital de la Timone à Marseille depuis 1994 et ayant un antécédent de maladie thrombo-embolique veineuse mais sans aucun des facteurs de risque principaux, c'est-à-dire une insuffisance en AntiThrombine (AT), en Protéine C (PC), en Protéine S (PS), une homozygotie pour la mutation FV Leiden (FVL) ou la mutation FII G20210A.

Tous les patients de l'étude MARTHA ont été génotypés soit avec la puce Illumina Human 610-Quad soit avec la puce Illumina Human 660W-Quad (Trégouët *et al.* 2009). Une imputation a ensuite été réalisée par Marine Germain (ingénieur de recherche au sein de l'équipe) pour déterminer les génotypes des polymorphismes non mesurés par les puces (Rocanin-Arjo *et al.* 2014) en utilisant le génome de référence du projet 1000 Genomes (T. 1. G. P. Consortium 2010). Un sous-groupe de 350 patients a été sélectionné aléatoirement parmi les cas de MTEV pour être épitypés, c'est-à-dire que l'on a mesuré le niveau de méthylation de l'ADN, à partir du sang périphérique, avec la puce Illumina Infinium HumanMethylation450k.

2 L'étude F5L-Pedigrees

L'étude F5L-Pedigrees a quant à elle été mise en place par le Dr. France Gagnon (Université de Toronto) à partir des patients du Dr. Phil Wells (Antoni *et al.* 2010). L'étude est composée de 255 sujets provenant de 5 grandes familles Franco-Canadiennes répartie sur 4 à 5 générations et sélectionnées à partir de patients ayant consulté avant 2005 la clinique de thrombophilie d'Ottawa pour une MTEV. Ces 5 grandes familles ont été sélectionnées à partir de cas idiopathiques (c'est-à-dire l'absence de facteur acquis tels qu'un cancer ou syndrome myéloprolifératif, d'une grossesse ou d'un post-partum, d'une immobilisation prolongée, d'un traumatisme, d'une chirurgie, d'un syndrome des anticorps antiphospholipides), sans mutation induisant un déficit en AT, en PS ou en PC, mais porteurs de la mutation FVL. L'objectif de cette étude était de diminuer l'hétérogénéité génétique entre les familles et ainsi augmenter la puissance des analyses de liaison.

Tous les sujets de l'étude F5L-Pedigrees ont également été génotypés avec la puce Illumina Human660W-Quad. Un sous-groupe de 227 sujets où l'ADN était encore disponible a été épitypé avec la puce Illumina Infinium HumanMethylation450k.

Chapitre 2

La puce HumanMethylation450k

Il est désormais possible de mesurer le niveau de méthylation de l'ensemble des gènes grâce à une technologie de biopuces (ou « micro-array »). La première puce à ADN pouvant mesurer la méthylation a été commercialisée en 2009 par la société Illumina sous le nom de Infinium HumanMethylation27k (HM27k) (Bibikova, Le *et al.* 2009), elle permettait de mesurer le niveau de méthylation d'environ 27 000 sites CpG à travers le génome. Fin 2011, une nouvelle génération de puce est apparue, il s'agit de la puce Infinium HumanMethylation450k (HM450k) (Bibikova, Barnes *et al.* 2011 ; Sandoval *et al.* 2011), qui cette fois mesure la méthylation sur environ 480 000 sites CpG ce qui correspond à environ 99% des gènes et ARN non codants connus.

Le fonctionnement de cette puce se base sur les propriétés du bisulfite de sodium. En effet, après une étape d'extraction de l'ADN des cellules, on traite l'ADN extrait au bisulfite de sodium qui va transformer les cytosines non méthylées en uracile tout en gardant intactes les cytosines méthylées. Ensuite, l'amplification du génome entier va convertir les uraciles obtenues en thymines. Pour résumer, à la fin de cette étape, les cytosines non méthylées d'un échantillon d'ADN deviendront des thymines alors que les cytosines méthylées resteront des cytosines. Les séquences d'ADN ainsi obtenues sont ensuite analysées via la puce HM450k sur laquelle se réalise une hybridation via une amorce allèle spécifique de 50 paires de bases et une extension d'une seule base marquée par un fluorochrome. Lors de cette étape, chaque séquence d'ADN va se fixer sur l'amorce qui aura la séquence complémentaire. La puce va permettre d'ainsi étudier la méthylation de 485 512 cytosines tout au long du génome.

1 Description des sondes

Les 485 577 sondes dont 485 512 mesurent la méthylation de cytosines sont réparties ainsi : 482 421 sur des sites CpG, 3 091 sur des sites CNG (N peut être un G, C, A ou T) et 65 sondes conçues pour mesurer non pas la méthylation mais la présence de SNPs (*Single-Nucleotide Polymorphism* ou polymorphisme d'un seul nucléotide) sélectionnés aléatoirement. Les 65 sondes mesurant un SNP ne sont pas conçues pour mesurer soit une cytosine soit une thymine (ancienne cytosine non méthylée convertie par le traitement au bisulfite de sodium) mais un polymorphisme connu. Par exemple une sonde mesure le SNP rs5936512, c'est-à-dire qu'elle détermine si l'échantillon d'ADN contient l'allèle ancestral (une guanine) ou l'allèle muté (une adénine). L'intérêt de ces 65 sondes est de pouvoir les utiliser pour le contrôle qualité notamment pour vérifier que les allèles détectés avec la puce HM450k correspondent bien à ceux détectés avec une autre méthode de génotypage.

Chaque sonde est composée d'une bille de silice d'une taille de 3 μm de diamètre fixée sur la puce et répliquée environ 15 fois. La puce est composée de deux types de sonde : 135 501 (27,9%) sondes sont de type Infinium I (utilisée notamment par l'ancienne génération de puce HM27k) et 350 076 (72,1%) sondes sont de type Infinium II (spécifique à la nouvelle puce HM450k).

Les sondes Infinium I sont des oligonucléotides de 50 bases. Elles mesurent l'état de méthylation d'un site CpG par paire, c'est-à-dire qu'une première sonde mesure l'état méthylé de la cytosine et une seconde mesure l'état non méthylé. La sonde mesurant l'état méthylé aura une séquence se terminant par une guanine pour s'hybriter avec la cytosine méthylée et donc non convertie en thymine. Alors que la sonde mesurant l'état non méthylé aura une séquence se terminant par une adénine pour s'hybriter avec la thymine résultant de la conversion de la cytosine non méthylée. Une fois la séquence d'ADN hybridée à la sonde ayant la séquence complémentaire en fonction de l'état de méthylation de la cytosine, se déroule alors une étape d'élongation de la sonde. L'élongation va permettre l'ajout d'une base azotée à la suite de l'adénine (sonde mesurant l'état non méthylé) ou de la guanine (sonde mesurant l'état méthylé). C'est cette base azotée (définie via le génome de référence lors de la conception de la puce HM450k) en position adjacente au site CpG, qui va émettre de la fluorescence par la libération d'un fluorochrome. La fluorescence sera rouge (émission d'une cyanine Cy5) lorsque la base azotée adjacente au site CpG sera une adénine ou une thymine. Tandis qu'elle sera verte (émission d'une cyanine Cy3) lorsque la base azotée adjacente au site CpG sera une cytosine ou une guanine. Les deux états d'un site CpG seront donc mesurés avec le même fluorochrome (figure 2.1)

car dépendant de la base azotée adjacente au site CpG.

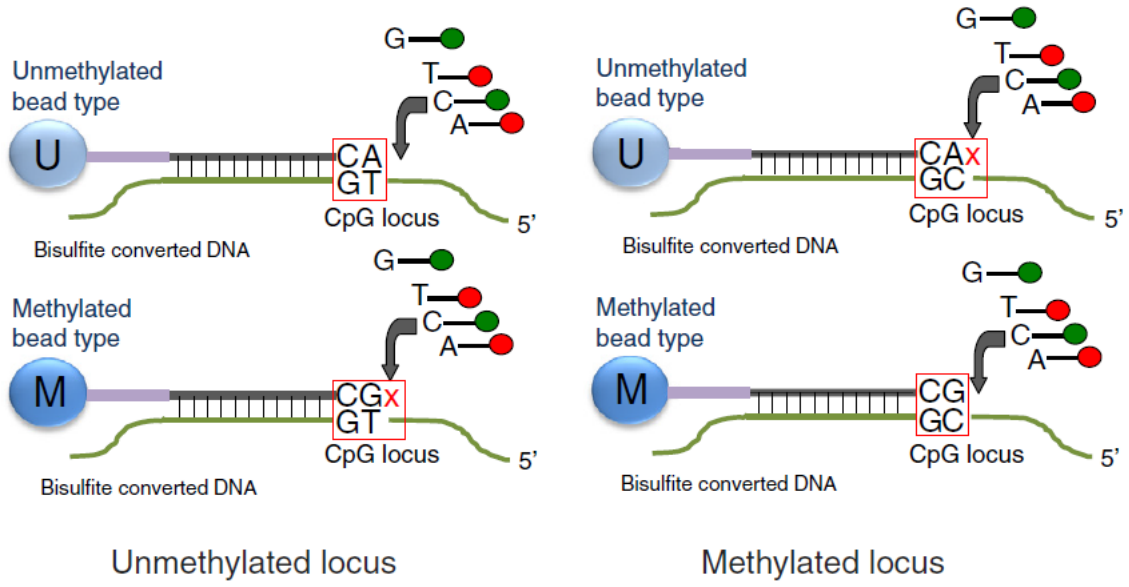

FIGURE 2.1 – Fonctionnement des sondes Infinium I. Schéma de BIBIKOVA, BARNES *et al.* 2011.

Les sondes Infinium II sont des oligonucléotides de 50 bases auxquels viennent se fixer une adénine avec un fluorochrome rouge (Cy5) si la cytosine complémentaire était non méthylée ou une guanine avec un fluorochrome vert (Cy3) si la cytosine complémentaire était méthylée. Pour un même site CpG, les sondes vont donc être mesurées dans le rouge et dans le vert en fonction de l'état méthylé de la cytosine (figure 2.2). Les sondes Infinium II permettent un gain de place sur la puce augmentant ainsi le nombre de sondes différentes et donc de loci étudiés.

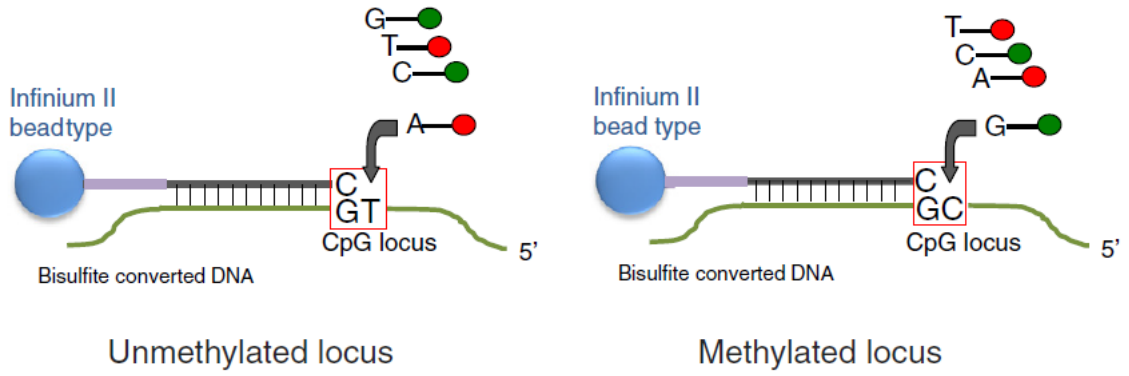

FIGURE 2.2 – Fonctionnement des sondes Infinium II. Schéma de BIBIKOVA, BARNES *et al.* 2011.

Pour résumer, la différenciation entre l'état méthylé de l'état non méthylé pour les sondes Infinium I se fera en fonction de la position sur la puce de la fluorescence. Tandis que pour les sondes Infinium II, la différenciation se fera en fonction de la couleur de la fluorescence émise.

2 Intensité de fluorescence

À partir de la puce, on mesure des intensités de fluorescence émises par le fluorochrome pour chaque état d'un site CpG, une pour l'état méthylé et une autre pour l'état non méthylé.

Sur la figure 2.3 est représentée par une courbe noire continue la fonction de densité des intensités des sondes mesurant l'état méthylé des sites CpG. Cette densité (en noire) peut être décomposée en fonction du type de sonde utilisée, Infinium I (courbe bleue continue) ou Infinium II (courbe rouge continue). Il en est de même pour les sondes mesurant l'état non méthylé des sites CpG qui sont représentées par une courbe noire en pointillé, également décomposable pour les sondes Infinium I (courbe bleue en pointillé) et pour les sondes Infinium II (courbe rouge en pointillé). Les intensités sont distribuées selon une loi Gamma (P. Du *et al.* 2010).

Apparaît également sur ce graphique la fonction de densité verte des intensités des sondes de contrôles négatifs émises dans le vert (Cy3) et la fonction de densité jaune des intensités des sondes de contrôles négatifs émises dans le rouge (Cy5). Le principe et le fonctionnement des sondes de contrôles négatifs sont décrits à la page 32 de ce document.

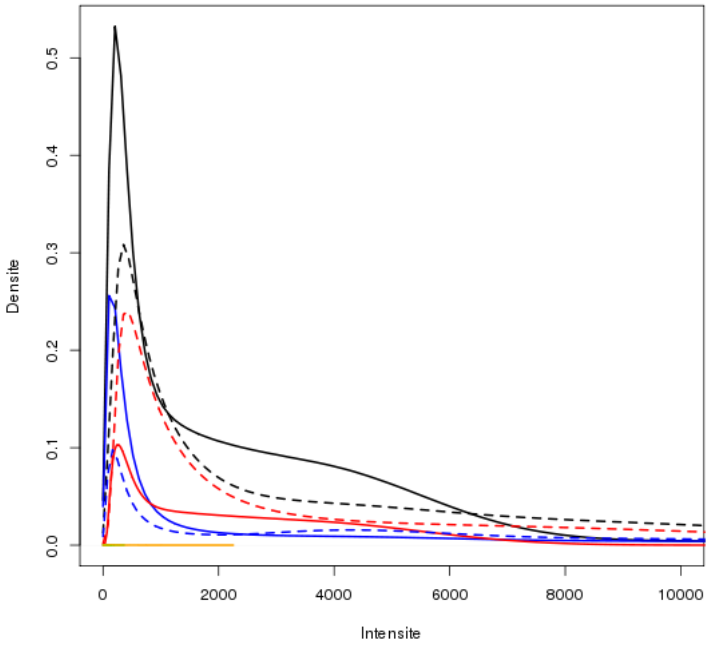

FIGURE 2.3 – Fonction de densité des intensités obtenue à partir des 588 échantillons analysés (MARTHA et F5L-Pedigrees).

2.1 Valeur β

À partir des intensités des sondes de mesure, on calcule pour chaque site CpG une valeur bêta (β) caractérisant son niveau de méthylation. Le β est calculé à partir de la formule suivante :

$$\beta = \frac{I_{meth}}{I_{meth} + I_{unmeth} + \alpha}$$

— I_{meth} : Intensité du signal pour la forme méthylée

— I_{unmeth} : Intensité du signal pour la forme non méthylée

— α : Compensation (*offset* en anglais) de 100 conseillé par Illumina

Les valeurs β comme leur nom l'indique sont distribuées selon une loi bêta. Il s'agit donc d'une valeur continue comprise entre 0 pour un site CpG non méthylé et 1 pour un site CpG méthylé. Cette valeur représente le pourcentage de méthylation du site CpG dans les cellules étudiées. Bien que l'état de méthylation pour une cytosine d'une molécule d'ADN donnée soit un état binaire, soit la cytosine est méthylée soit elle ne l'est pas, ici la valeur β obtenue provient des intensités mesurées sur plusieurs cellules et potentiellement plusieurs types cellulaires comme c'est le cas lorsque l'ADN provient du sang périphérique.

La figure 2.4 nous montre la fonction de densité des β obtenus pour les 588 échantillons (MARTHA et F5L-Pedigrees). On distincte clairement deux pics, un pour l'état non méthylé à gauche et le second pour l'état méthylé à droite. On voit cependant que le pic de droite possède un second plus petit pic qui lui est juxtaposé. On verra dans le chapitre détaillant les biais (p. 48) qu'il s'agit d'un biais provoqué par l'utilisation de deux types de sondes.

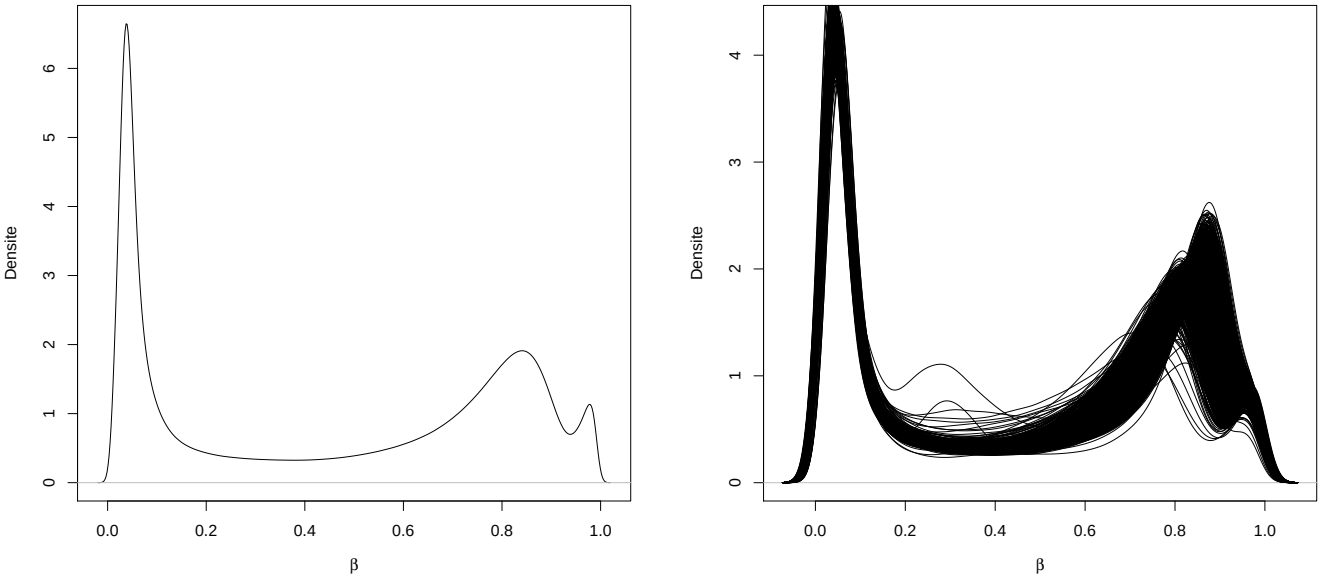

FIGURE 2.4 – Fonction de densité des β obtenue à partir des 588 échantillons analysés. Fonction globale à gauche et fonctions individuelles à droite.

2.2 Valeur M

Les valeurs β sont intuitives ce qui rend les interprétations biologiques plus faciles, en effet il s'agit du pourcentage de méthylation du site CpG. Mais comme elles sont majoritairement comprises dans l'intervalle de 0 et 0,2 pour l'état non méthylé ainsi que dans l'intervalle de 0,8 et 1 pour l'état méthylé, elles sont donc soumises à de l'hétéroscédasticité. Ce problème d'hétéroscédasticité empêche l'utilisation de nombreux modèles statistiques qui imposent une homoscédasticité dans les données. Une alternative à la valeur β, la valeur M a donc été proposée par P. DU *et al.* 2010. La transformation en valeurs M permet de stabiliser la variance et ainsi de supprimer le problème d'hétéroscédasticité présent avec les valeurs β. Les valeurs M ont ainsi des meilleures propriétés statistiques mais sont plus difficiles à interpréter biologiquement. Pour obtenir des valeurs M, on transforme les valeurs β via la fonction *logit* :

$$M = log_2\left(\frac{\beta}{1-\beta}\right) = log_2\left(\frac{I_{meth}}{I_{unmeth}+\alpha}\right)$$

Une variante de la définition de la valeur M peut également être utilisée, le compensateur α est ajouté à la fois à l'intensité de l'état méthylé et à l'intensité de l'état non méthylé :

$$M = log_2\left(\frac{I_{meth}+\alpha}{I_{unmeth}+\alpha}\right)$$

Une des conséquences de cette transformation est de ne plus avoir un intervalle compris entre 0 et 1 comme c'est le cas pour les valeurs β mais un intervalle compris entre $-\infty$ et $+\infty$. La figure 2.5 nous montre la fonction de densité des M obtenus pour les 588 échantillons.

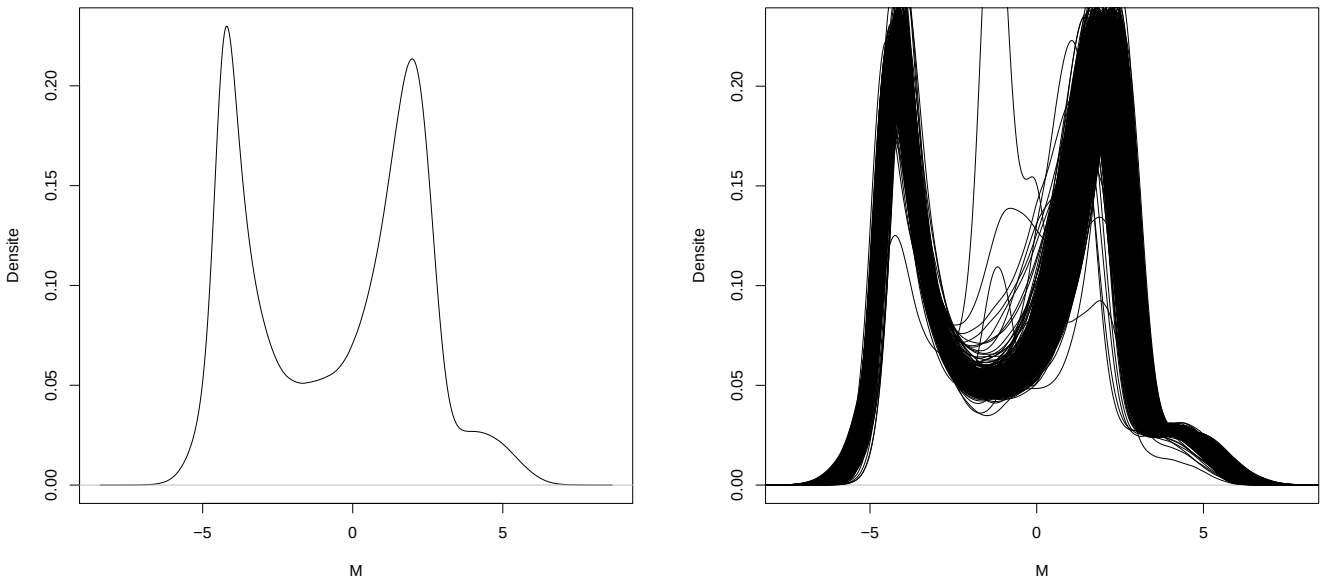

FIGURE 2.5 – Fonction de densité des M obtenue à partir des 588 échantillons analysés. Fonction globale à gauche et fonctions individuelles à droite.

Bien que les valeurs M soient plus efficaces que les valeurs β en terme de puissance et de précision pour identifier des sites CpG différemment méthylés lorsque l'on compare deux populations (P. DU *et al.* 2010), dans certains cas l'utilisation des valeurs β reste plus efficace (WAHL *et al.* 2014 ; ZHUANG *et al.* 2012). En effet, bien que la distribution globale des valeurs β soit bimodale, lorsque l'on regarde la distribution des valeurs β d'un unique site CpG, celle-ci suit le plus souvent une loi normale et peut donc être utiliser dans les modèles statistiques couramment employés.

Chapitre 3

Contrôle Qualité et Normalisation

Les données issues de l'utilisation de biopuces sont soumises à de nombreux biais expérimentaux (Sabbah *et al.* 2011 ; Siegmund 2011), certains communs à toute les puces à ADN et d'autres spécifiques à la puce HM450k. Avant de pouvoir étudier par des méthodes bio-informatiques et biostatistiques si des phénotypes biologiques et/ou cliniques peuvent être associés à des profils de méthylation particuliers, il est important en premier lieu de vérifier la présence des biais potentiels, puis de les corriger afin de limiter leur impact lors de l'analyse plus fine des données de méthylation en relation avec des phénotypes d'intérêt.

Au début de ce travail de thèse, la puce HM450k était disponible depuis peu de temps, tous les biais n'étaient pas encore connus et correctement corrigés par des méthodes statistiques. La littérature proposait quelques méthodes pour tenter de corriger les biais, mais aucune ne se dégageait des autres. J'ai dû identifier les biais présents dans mes données et proposer, adapter, comparer différentes méthodes pour les corriger avant de pouvoir appliquer différentes approches statistiques (Kuan *et al.* 2010 ; Siegmund 2011 ; Siegmund & Laird 2002) pour identifier des signatures de méthylation associées aux variables biologiques d'intérêt. La partie suivante aborde le contrôle qualité ainsi que les biais et la normalisation des données obtenues par la puce HM450k. Les données de méthylation des études MARTHA et F5L-Pedigrees ont été générées au même moment au TCGA de Toronto, il était donc plus judicieux de réaliser le contrôle qualité et la normalisation en même temps pour les deux études.

1 Sondes de contrôles

La puce inclut en plus des 485 577 sondes mesurant soit des niveaux de méthylation soit l'allèle d'un polymorphisme, 850 sondes de contrôles permettant d'évaluer différents critères de performances du déroulement des différentes étapes de la mesure du méthylome via la puce. Parmi les 850 sondes de contrôles, certaines évaluent la performance d'une étape précise lors du processus tandis que d'autres évaluent la performance entre les échantillons. Ces 850 sondes de contrôles sont réparties ainsi :

— 6 sondes de *contrôles de coloration* : permettant d'examiner l'efficacité de l'étape de coloration dans les deux canaux (rouge et vert) indépendamment de l'étape d'hybridation et d'extension.

— 4 sondes de *contrôles d'extension* : conçues pour tester l'efficacité d'extension de nucléotides T, A, C et G, sur une sonde en épingle à cheveux. L'extension est réalisée par une polymérase qui va synthétiser le brin complémentaire à la séquence d'ADN à la suite de la sonde. Dans ce contexte, l'extension ne se fera que d'un seul nucléotide.

— 3 sondes de *contrôles d'hybridation* : permettant de tester la performance globale de la puce HM450k en utilisant des cibles synthétiques à la place de l'ADN amplifié. Ces cibles synthétiques sont parfaitement complémentaires à la séquence de la sonde.

— 2 sondes de *contrôles de suppression de cibles* : conçues pour contrôler l'étape de nettoyage suivant la réaction d'extension. Les sondes sont conçues de telle sorte que l'extension ne se produise pas. Ces contrôles doivent aboutir à un signal plus faible que celui des contrôles d'hybridation, ce qui indique que les cibles ont été efficacement éliminées après l'extension. En effet, une fois l'hybridation puis l'extension réalisée, il est nécessaire de "nettoyer" les puces pour éliminer tous les éléments non fixés et ainsi mesurer la fluorescence des fluorochromes fixés.

— 12 sondes de *contrôles de conversion au bisulfite I* : permettant d'évaluer l'efficacité de la conversion au bisulfite de l'ADN génomique. Ces sondes interrogent un polymorphisme C/T créé par conversion au bisulfite de sodium. Si la conversion au bisulfite est réussie, les sondes "C" correspondant à la séquence convertie seront prolongées. Si la conversion a échoué, les sondes "U" correspondant à la séquence non convertie seront prolongées.

— 4 sondes de *contrôles de conversion au bisulfite II* : idem mais avec une conception de type Infinium II.

— 12 sondes de *contrôles de spécificité I* : conçues pour surveiller le potentiel d'extension non spécifique des sondes en ciblant des sites T non polymorphes.

— 3 sondes de *contrôles de spécificité II* : idem mais avec une conception de type Infinium II.

— 614 sondes de *contrôles négatifs* : conçues pour ne pas avoir de séquence complémentaire parmi les fragments d'ADN et par conséquent ne pouvant pas s'hybrider. En l'absence d'hybridation, aucun fluorochrome ne peut se fixer et l'intensité détectée pour ces sondes devrait donc être nulle. On détecte toutefois un signal pour ces sondes, ce qui permet d'estimer le bruit de fond (voir le chapitre détaillant le biais du bruit de fond, p. 46).

— 4 sondes de *contrôles non polymorphiques* : permettant de tester la performance globale de l'analyse, de l'amplification de la détection, en interrogeant une base particulière dans une zone non polymorphe du génome. Ces sondes permettent de comparer les performances de la puce entre différents échantillons.

— 32 sondes de *contrôles de normalisation Adénine* : conçues pour cibler la même région dans les gènes de ménage et ne comprennent pas de sites CpG. L'extension se fait par l'ajout d'une adénine.

— 61 sondes de *contrôles de normalisation Cytosine* : idem mais avec l'ajout d'une cytosine.

— 32 sondes de *contrôles de normalisation Guanine* : idem mais avec l'ajout d'une guanine.

— 61 sondes de *contrôles de normalisation Thymine* : idem mais avec l'ajout d'une thymine.

Il est nécessaire de contrôler les intensités obtenues par toutes ces sondes pour vérifier les performances des différentes étapes et ainsi pouvoir exclure les échantillons pour lesquels il y a eu des défaillances lors du processus. Par exemple sur la figure 3.1 qui représente les diagrammes en boîte des intensités obtenues par les sondes de contrôles de conversion au bisulfite I pour les 588 échantillons mesurés, on peut observer que les intensités des sondes "U" correspondant à la séquence non convertie sont faibles, ce qui indique que les séquences ont bien été converties. Cela est confirmé par les intensités des sondes "C" correspondant à la séquence convertie qui ont des valeurs plus fortes et donc que l'étape de conversion a bien été réalisée.

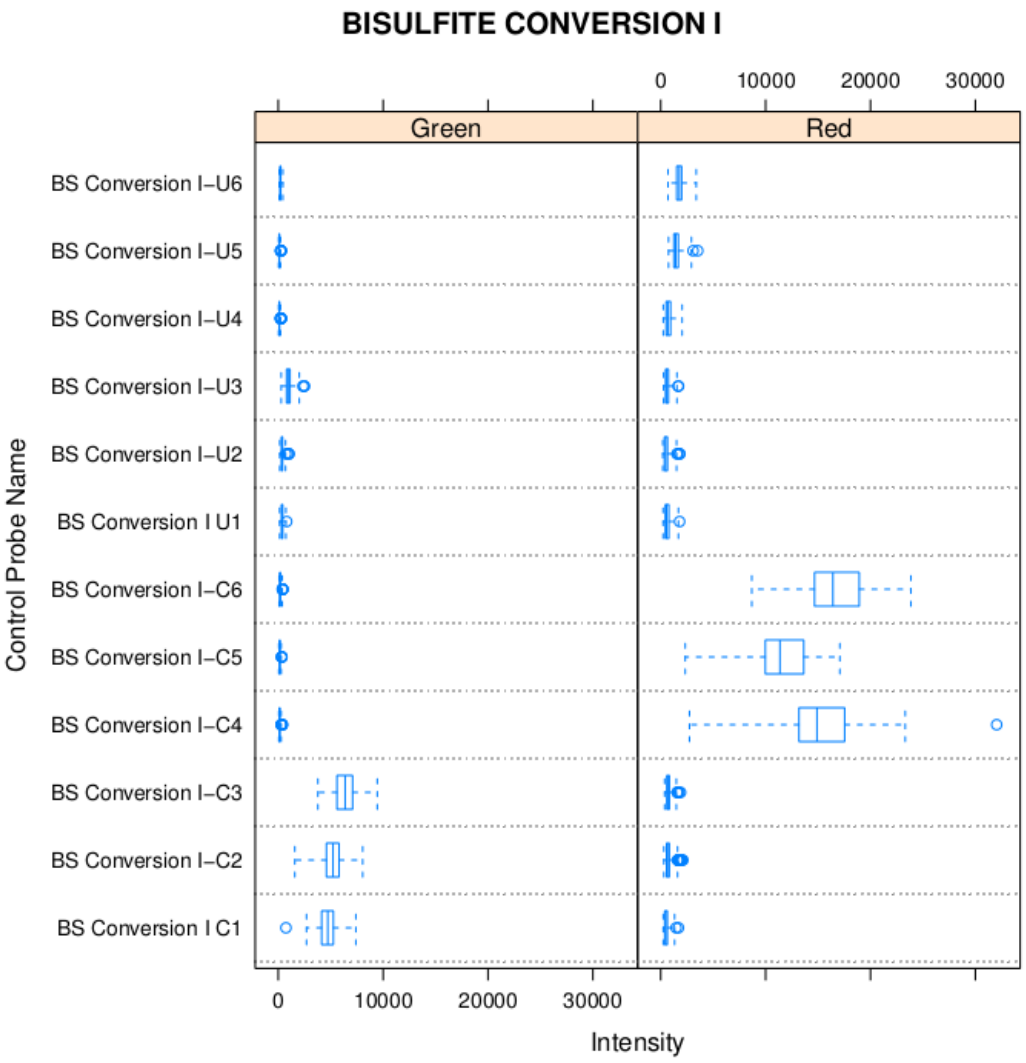

FIGURE 3.1 – Diagrammes en boîte des intensités obtenues par les 12 sondes de contrôles de conversion au bisulfite I pour les 588 échantillons mesurés. Dans le vert à gauche et dans le rouge à droite.

Un autre exemple avec la figure 3.2 qui représente les diagrammes en boîte des intensités obtenues par les sondes de contrôles d'extension pour les 588 échantillons mesurés. On peut y voir que les intensités des extensions via une adénine ou une thymine sont fortes dans la fluorescence rouge et qu'il n'y a pas de fluorescence verte alors que nous avons l'inverse pour les extensions via des cytosines ou des guanines. Ce graphique montre que les extensions ont bien été réalisées et on peut observer également que les intensités obtenues par les deux fluorochromes ne sont pas équivalentes, le fluorochrome vert (Cy3) a une intensité moyenne comprise entre 10000 et 15000 alors qu'elle est comprise entre 30000 et 40000 pour le fluorochrome rouge (Cy5). Cela est dû à des propriétés physico-chimiques différentes entre les deux fluorochromes qui conduisent à des efficacités d'hybridation différentes et à l'introduction d'un biais décrit à la page 50.

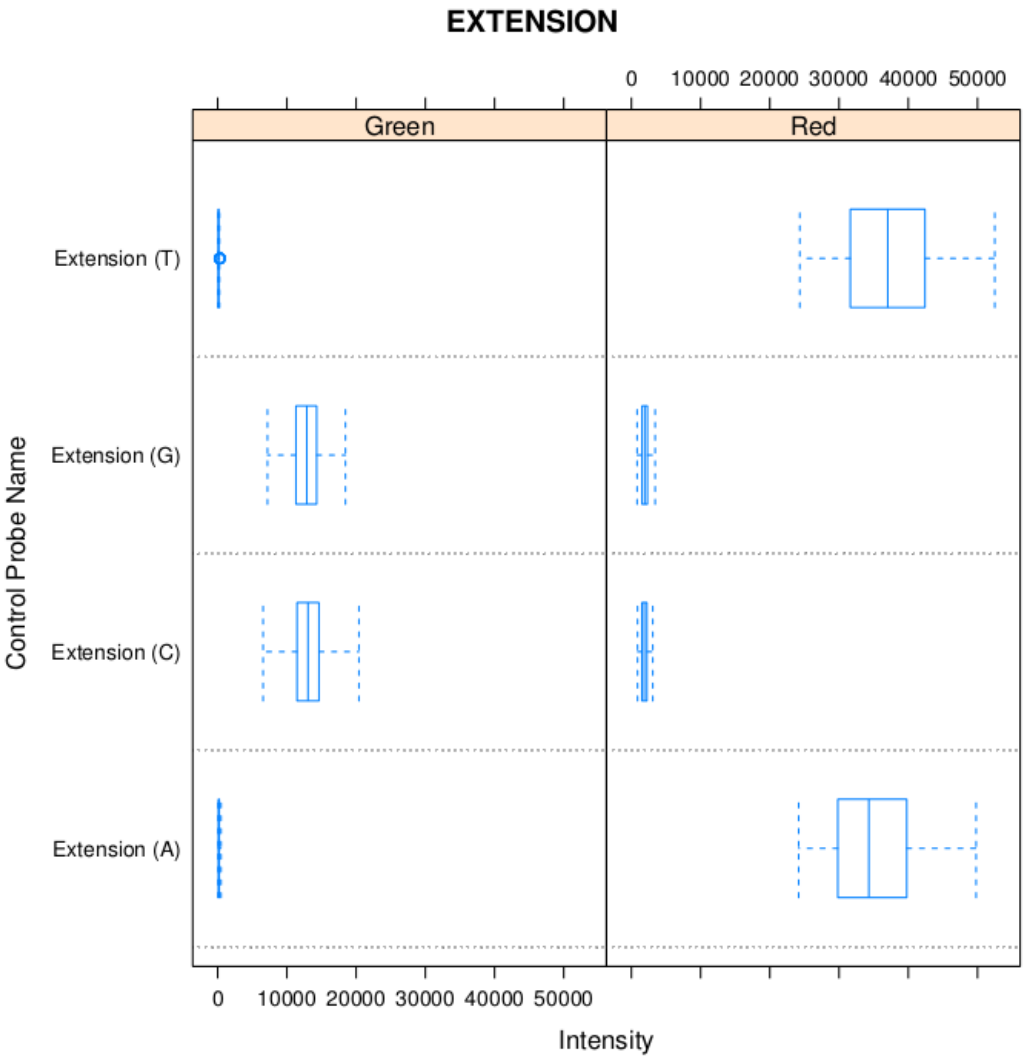

FIGURE 3.2 – Diagrammes en boîte des intensités obtenues par les sondes de contrôles d'extension pour les 588 échantillons mesurés.

En ce qui concerne les sondes de contrôles négatifs qui permettent d'estimer le bruit de fond, leur fonctionnement et leur intérêt sont expliqués à la page 46. Les sondes de contrôles négatifs étant peu nombreuses par rapport aux sondes mesurant les niveaux de méthylation, on ne distingue pas correctement leurs fonctions de densité sur le graphique de la figure 2.3. Pour mieux représenter leurs fonctions de densité, un zoom a été réalisé sur ces sondes (figure 3.3). Cette figure nous montre que l'intensité de la fluorescence émise dans le rouge (Cy5) est plus importante que l'intensité de la fluorescence émise dans le vert (Cy3).

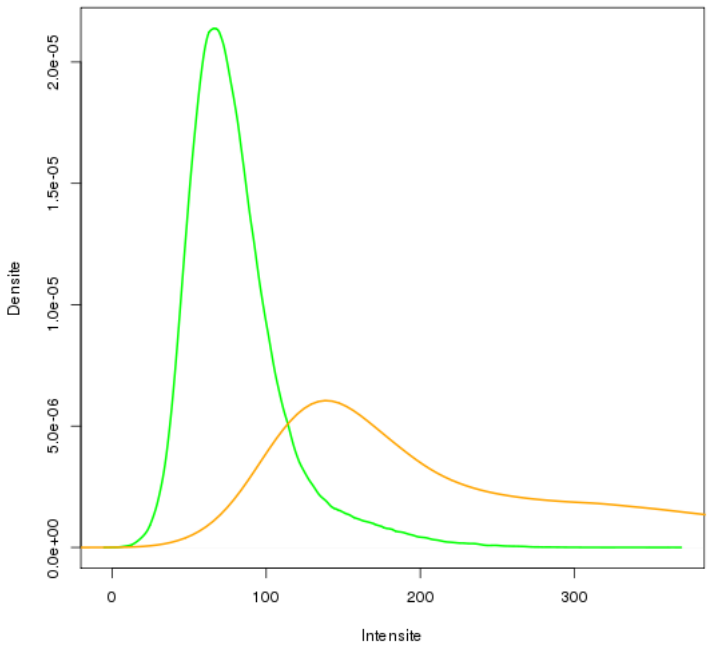

FIGURE 3.3 – Fonction de densité des intensités des sondes de contrôles négatifs obtenue à partir des 588 échantillons analysés. En vert, la fonction de densité des intensités des contrôles négatifs émises dans le vert (Cy3) et en jaune, la fonction de densité des intensités des contrôles négatifs émises dans le rouge (Cy5).

2 Valeur p de détection

À chaque valeur β est associée une valeur p de détection (*detection p-value* en anglais). Elle reflète la force de l'intensité du signal d'hybridation de l'ADN par rapport à l'intensité du bruit de fond, c'est-à-dire la puissance de l'intensité des sondes de mesures par rapport à la puissance de l'intensité des sondes de contrôles négatives. Elle peut être utilisée pour mesurer objectivement la performance globale de la sonde. Une valeur p non significative est généralement la conséquence d'une mauvaise conception de la sonde, d'une mauvaise hybridation ou d'une anomalie chromosomique dans l'ADN étudié (ex : mutations ou indels) sur l'emplacement correspondant à la sonde. Illumina conseille d'exclure les sondes ayant une valeur p de détection supérieure à 0,05 pour plus de 5% des échantillons mesurés.

À ce jour, deux méthodes sont communément utilisées pour calculer la valeur p de détection. La première méthode utilisée par Illumina dans le logiciel GenomeStudio ainsi que par le package "*methylumi*" utilise la fonction de répartition empirique. La seconde méthode est utilisée quant à elle dans le package "*minfi*". L'efficacité d'une sonde étant dépendante de l'échantillon, la valeur p de détection est à calculer pour chaque site CpG et pour chaque échantillon indépendamment.

2.1 Méthode de calcul selon la méthode "methylumi"

La valeur p de détection est calculée par site CpG et par échantillon dans le package "*methylumi*" selon la méthode suivante : On calcule la fonction de répartition empirique des sondes de contrôles négatifs par individu et par fluorochrome (Cy3 et Cy5). Grâce à la fonction de répartition empirique, on en déduit le centile correspondant à l'intensité des sondes de mesures en tenant compte du fluorochrome utilisé. On obtient par sondes deux centiles correspondant aux intensités pour l'état méthylé (noté centileM) et pour l'état non méthylé (centileUnM). La valeur p de détection sera la valeur minimale entre $1 - centileM$ et $1 - centileUnM$.

2.2 Méthode de calcul selon la méthode "minfi"

La valeur p de détection est calculée par site CpG et par échantillon dans le package "*minfi*" selon la méthode dite "$m + u$". Pour cela on détermine la probabilité q qu'une valeur soit inférieure à l'intensité totale (somme des intensités de l'état méthylé et de l'état non méthylé) de la sonde à partir d'une loi normale $N \sim (\mu, \sigma^2)$.

En fonction du type de sonde pour lequel on calcule la valeur p de détection, μ prendra comme valeur :

— Soit $\mu = MED_{rouge} * 2$ dans le cas d'une sonde de type Infinium I et mesurée dans le rouge ;

— Soit $\mu = MED_{vert} * 2$ dans le cas d'une sonde de type Infinium I et mesurée dans le vert ;

— Soit $\mu = MED_{rouge} + MED_{vert}$ dans le cas d'une sonde de type Infinium II.

Avec MED la médiane de l'intensité de fluorescence du bruit de fond, indépendamment pour le vert et le rouge (notés respectivement MED_{vert} et MED_{rouge}).

Il en est de même pour la valeur σ qui prendra comme valeur :

— Soit $\sigma = MAD_{rouge} * 2$ dans le cas d'une sonde de type Infinium I et mesurée dans le rouge ;

— Soit $\sigma = MAD_{vert} * 2$ dans le cas d'une sonde de type Infinium I et mesurée dans le vert ;

— Soit $\sigma = MAD_{rouge} + MAD_{vert}$ dans le cas d'une sonde de type Infinium II.

Avec MAD l'écart médian absolu (MAD pour *Median Absolute Deviation* anglais) de l'intensité de fluorescence du bruit de fond, indépendamment pour le vert et le rouge (notés respectivement MAD_{vert} et MAD_{rouge}).

Pour obtenir la valeur p de détection il suffira de prendre $p = 1 - q$.

2.3 Comparaison des valeurs p de détection

Nous avons comparé les deux méthodes de calcul de la valeur p de détection pour savoir laquelle était la plus pertinente. La comparaison a porté sur les 485 577 sondes et sur les 588 échantillons mesurés (MARTHA, F5L-Pedigrees et les 11 réplicats) soit sur un total de 285 519 276 sondes. Le calcul de la valeur p de détection a été réalisé pour les 285 519 276 sondes selon les deux méthodes ("methylumi" et "minfi").

Le tableau 3.1 indique le nombre de sondes exclues lorsque l'on calcule la valeur p de détection selon la méthode "methylumi" et que l'on fixe le seuil de rejet à 0,05 pour plus de 5% des échantillons.

TABLE 3.1 – Nombre de sondes exclues en calculant la valeur p de détection selon la méthode "methylumi".

Localisation Génomique	Nombre de sondes exclues	Nombre total de sondes	%
Autosomes	1027	473929	0.21
Chromosome Y	0	416	0
Chromosome X	80	11232	0.71
Total	1107	485577	0.23

La méthode "methylumi" va donc exclure 1107 sondes (soit 0,23% de la globalité des sondes) pour lesquelles les valeurs p de détection sont supérieures à 0,05 pour plus de 5% des individus. En calculant la valeur p de détection avec la méthode "minfi", on obtient le tableau 3.2.

TABLE 3.2 – Nombre de sondes exclues en calculant la valeur p de détection selon la méthode "minfi".

Localisation Génomique	Nombre de sondes exclues	Nombre total de sondes	%
Autosomes	5642	473929	1.15
Chromosome Y	13	416	3.16
Chromosome X	547	11232	4.87
Total	6202	485577	1.28

Avec la méthode "minfi" et le même seuil de rejet, on exclut cette fois 6202 sondes (soit 1,28%). La méthode "minfi" est donc plus stricte que la méthode "methylumi".

Bien que la méthode proposée par la Illumina soit de calculer la valeur p de détection selon la méthode "methylumi" et d'exclure les sondes pour lesquelles la valeur p de détection est supérieure à 0,05 pour plus de 5% des échantillons, nous avons choisi de garder le même seuil mais de sélectionner la méthode de calcul "minfi". Nous avons donc exclut les 6202 sondes pour lesquelles la valeur p de détection (estimée selon la méthode "minfi") était supérieure à 0,05 dans plus de 5% des 588 échantillons mesurés (MARTHA, F5L-Pedigrees et les 11 réplicats).

3 Sondes sur les chromosomes sexuels

Une des étapes du contrôle qualité est de vérifier que les sondes fixant une région du chromosome Y n'émettent pas de signal chez les individus du sexe féminin. On voit sur la figure 3.4 qui représente les fonctions de densité des valeurs β du chromosome Y en fonction du sexe des individus qu'il y a effectivement deux pics pour les hommes, ce qui correspond à un état non méthylé pour le pic proche de 0 et méthylé pour le pic proche de 1. Alors que pour les femmes qui ne possèdent pas de chromosome Y, le signal est dû à des valeurs β ayant une valeur p de détection supérieure au seuil de significativité (résultants de la non hybridation des sondes aux fragments d'ADN) ou à de la réactivité croisée (voir la p. 42).

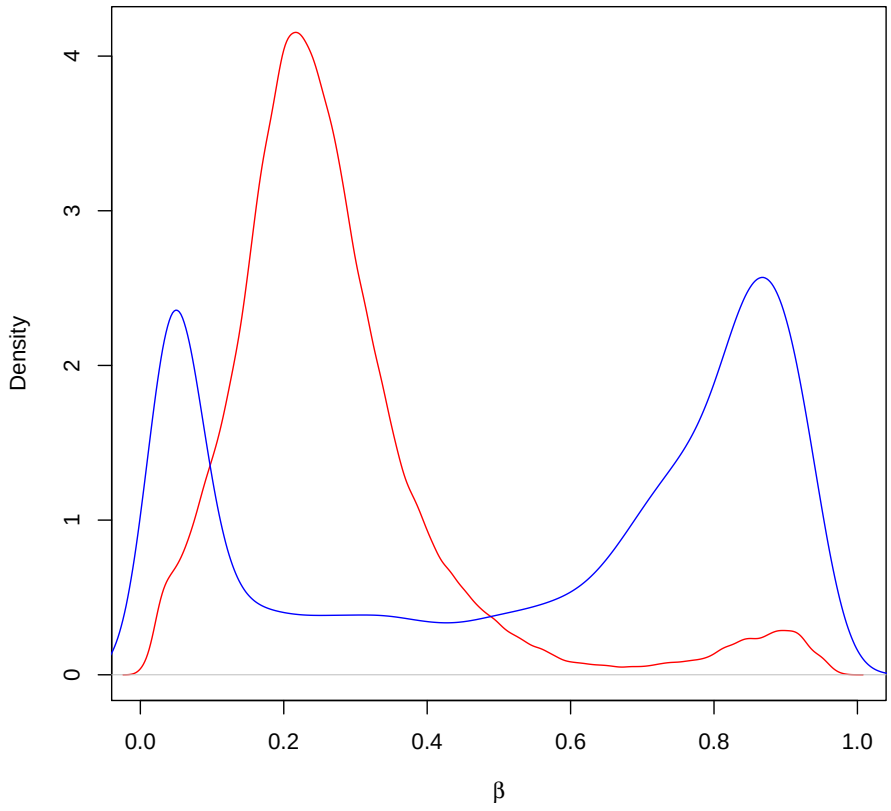

FIGURE 3.4 – Fonction de densité des valeurs β du chromosome Y en fonction du sexe des individus (rouge pour Femme et bleue pour Homme) chez les 588 échantillons.

La figure 3.5 qui représente des diagrammes en boîte des intensités des 511 sondes se fixant sur le chromosome Y chez 50 femmes dans l'étude MARTHA. On peut voir que pour deux d'entre elles les

intensités sont plus importantes, le 3^{me} quartile est aux alentours de 3000 alors qu'il est proche de 0 pour les autres femmes. On peut donc facilement en déduire que ces deux échantillons possédaient un chromosome Y et qu'il ne s'agissait pas de femmes mais d'hommes. Il y a probablement eu une erreur au cours des différentes étapes de collecte des données inversant probablement des identifiants.

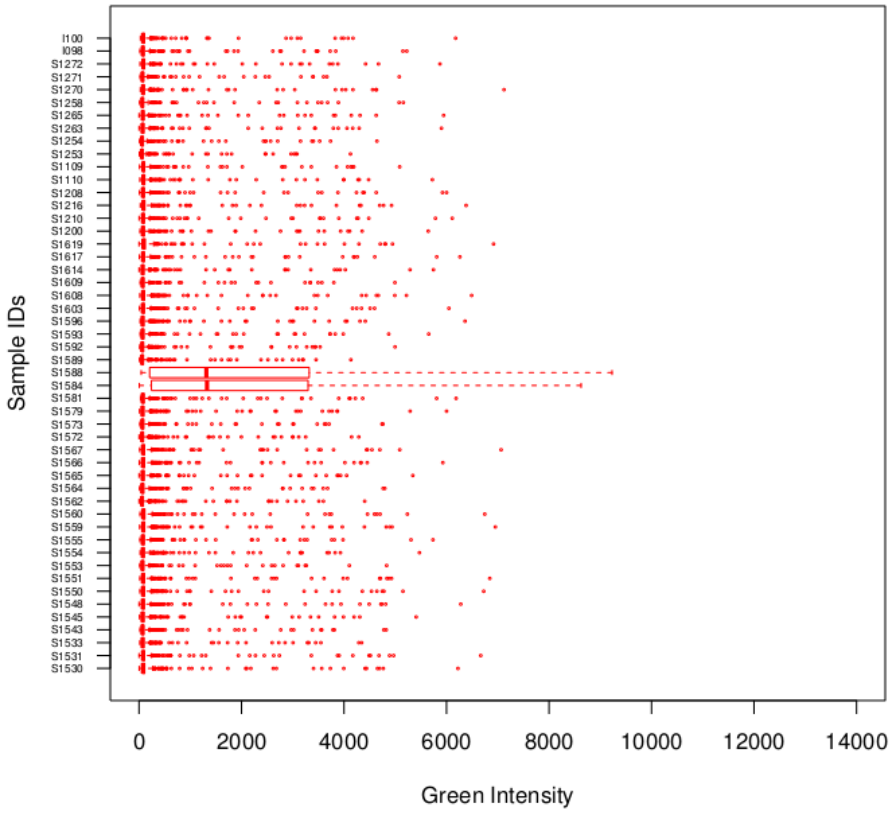

FIGURE 3.5 – Diagrammes en boîte des intensités (en abscisse) des 511 sondes se fixant sur le chromosome Y chez 50 femmes (en ordonnée) dans l'étude MARTHA.

Cette vérification permet d'exclure les échantillons pour lesquels il y a une incohérence entre les sexes phénotypiques et les données génétiques. Les sondes se fixant sur le chromosome X ne sont pas totalement adaptées pour la vérification du sexe car les profils de méthylation ne différent pas nécessairement entre les hommes et les femmes. Par exemple, lorsque toutes les molécules d'ADN à un site CpG donné sur l'unique chromosome X d'un homme sont méthylées, alors la valeur β sera proche de 1. Il est toutefois possible de réaliser la vérification sur le chromosome X mais les différences seront moins importantes que lorsque l'on réalise la vérification sur le chromosome Y.

4 Sites CpG polymorphes

La présence au sein de la sonde d'un polymorphisme nucléotidique (SNP) peut biaiser la mesure du β (Y.-a. Chen *et al.* 2013). En effet, la présence d'un SNP va entraîner des modifications de la séquence qui vont se répercuter par des modifications de l'affinité entre la séquence d'ADN et la sonde et donc sur une hybridation moins efficace. Plus le polymorphisme sera proche du site CpG, plus son impact sur la valeur β sera important. Cela est d'autant plus vrai lorsque le polymorphisme sera de type C/T, car dans ce cas on ne pourra distinguer si on mesure bien l'allèle T dû à la conversion au bisulfite ou s'il est antérieur au traitement. Il est donc important d'exclure les sondes ayant des polymorphismes au niveau du site CpG. Pour cela, il est nécessaire de mettre à jour les annotations d'Illumina avec par exemple les données provenant du projet 1000 Genomes (T. 1. G. P. Consortium 2010) pour pouvoir filtrer les sondes ayant des SNPs non connus lors de la mise sur le marché de la puce HM450k. À partir des données de Y.-a. Chen *et al.* 2013, j'ai exclu 66877 sondes pour lesquelles il pouvait potentiellement y avoir un SNP sur le site CpG quel que soit la fréquence du polymorphisme. Une autre possibilité si le génotype des individus est connu est de vérifier la présence ou non du SNP chez les individus et exclure ceux qui possèdent le polymorphisme.

La figure 3.6 illustre le cas d'un site CpG (cg19949776) influencé par un SNP (rs8031702) et où l'on ne peut pas facilement identifier si la variation du β est dû à la méthylation ou au SNP.

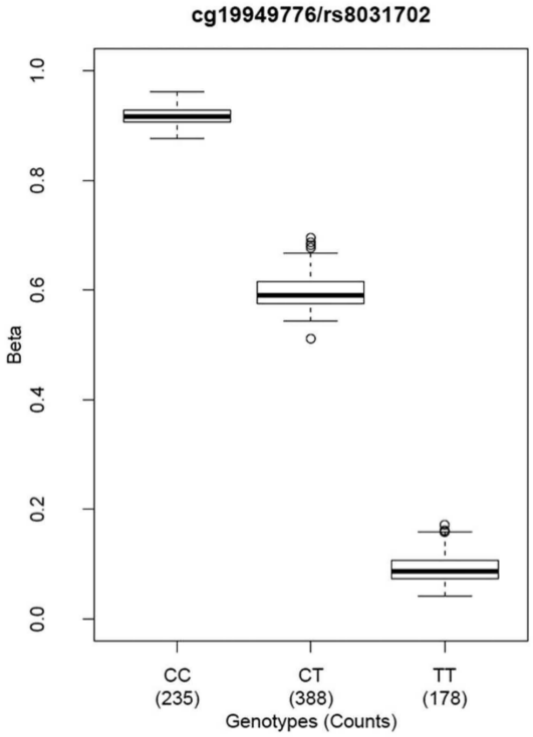

FIGURE 3.6 – Diagrammes en boîte des β du site CpG cg19949776 en fonction des génotypes du polymorphisme rs8031702. Extrait de Y.-a. CHEN *et al.* 2013.

Dans le cas du site CpG cg19949776, il existe un polymorphisme C/T (rs8031702) au niveau de la cytosine du site CpG. L'allèle C est détecté comme un allèle méthylé tandis que l'allèle T est détecté comme un allèle non méthylé. Dans ce cas de figure, il n'est pas aisé de savoir si la valeur β représente le niveau de méthylation ou si elle est biaisé par la présence du polymorphisme.

5 Réactivité croisée

Certaines sondes présentent une réactivité croisée (Y.-a. CHEN *et al.* 2013 ; X. ZHANG *et al.* 2012), c'est-à-dire que la sonde n'est pas spécifique à une seule séquence du génome mais est également à une autre (voir plus) séquence du génome. La sonde va donc mesurer la méthylation à plusieurs loci du génome, ce qui nous empêche de pouvoir interpréter correctement le signal obtenu. Les 30969 sondes possédant une réactivité croisée ont donc été exclues de l'analyse.

6 Les différents filtrages des données

Les échantillons et les sondes ayant des données aberrantes doivent être exclus des futures analyses.

Une Analyse en Composantes Principales (ACP ou *Principal Component Analysis* en anglais) à été réalisée sur les valeurs β non corrigées des 588 échantillons. L'ACP est une méthode d'analyse de données très utilisée dans le domaine de la recherche biomédicale notamment en ce qui concerne l'analyse des données "omiques". L'objectif de cette méthode est de réduire le nombre de dimensions des données en calculant les composantes (ou dimensions) qui expliquent la plus forte proportion de la variabilité des données. La projection des données dans un plus petit espace permet de réduire le nombre de variables. L'ACP va donc créer un axe (ou composante) expliquant la plus forte proportion de la variabilité des données. Puis ensuite, elle va créer un nouvel axe orthogonal au premier (c'est-à-dire non corrélé) expliquant une nouvelle fois la plus forte proportion de la variabilité. Par la suite, l'ACP va ainsi créer de nouveaux axes orthogonaux aux précédents et de variance maximale. Les échantillons pourront par la suite être caractérisés à partir de leurs coordonnées sur les axes ainsi créés. L'utilisation de l'ACP a permis ici d'identifier et d'exclure 4 échantillons ayant des valeurs aberrantes. Ces 4 échantillons avaient des valeurs pour les deux premières composantes trop éloignées des autres échantillons. En projetant les individus sur les deux premières dimensions créées par l'ACP à partir des valeurs β brutes, ces 4 individus se retrouvaient éloignés des autres individus. La figure 3.7 représente les échantillons des études MARTHA (en vert) et F5L-Pedigrees (en bleu) projetés sur les trois premières composantes.

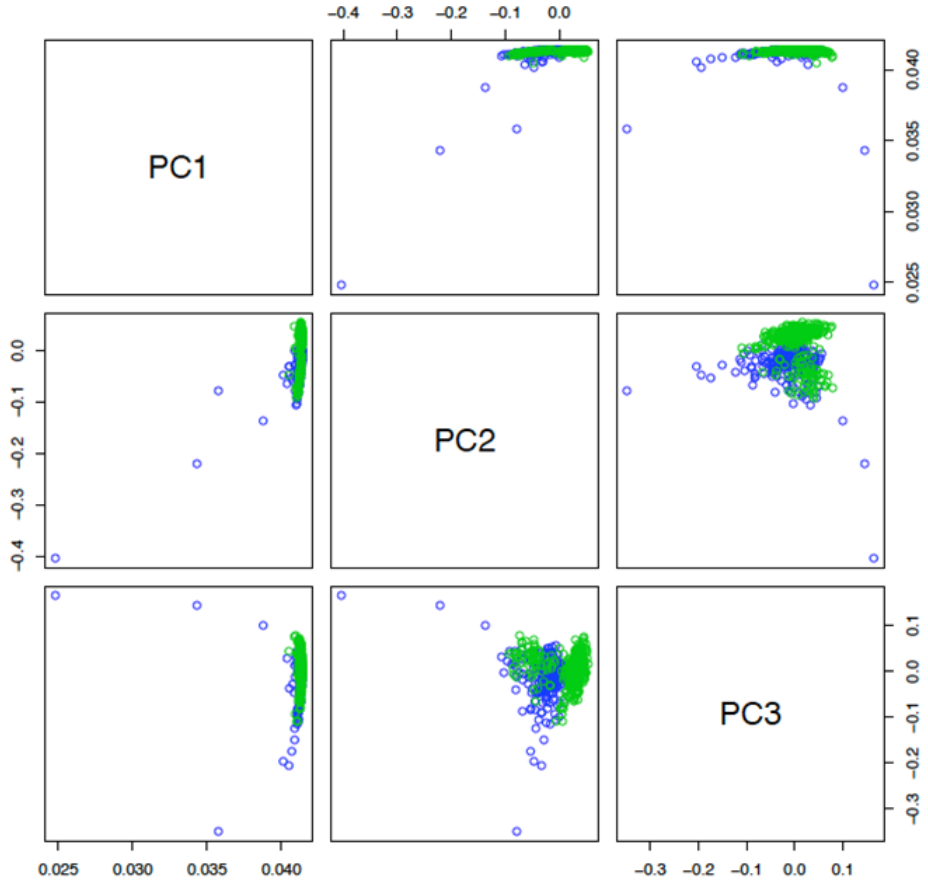

FIGURE 3.7 – Représentation graphique des échantillones des études MARTHA (en vert) et F5L-Pedigrees (en bleu) en fonction des trois premières composantes.

Par ailleurs, ces 4 échantillons avaient également de mauvais résultats aux sondes de contrôles (voir la page 32 qui détaille les sondes de contrôles) ce qui conforte le choix de les exclure des futures analyses statistiques. J'ai ensuite exclu 11 échantillons qui étaient des réplicats. Pour sélectionner laquelle des deux mesures des réplicats allait être exclus, j'ai comparé visuelle les résultats des sondes de contrôles. Cela m'a permis d'exclure la mesure pour laquelle les sondes de contrôles avaient les résultats les plus mauvais. Au final, les analyses décrites dans ce travail de thèse ont été réalisées sur 350 individus de l'étude MARTHA et 223 individus de l'étude F5L-Pedigrees.

Pour l'ensemble des échantillons, j'ai exclu sur les 485577 sondes de la puce HM450k celles ayant une réactivé croisé ($n = 30969$), celles ayant un polymorphisme sur le site CpG ($n = 66877$), celles ayant une valeur p de détection trop élevée ($n = 6202$) ainsi que celles conçues pour mesurer un polymorphisme et non la méthylation ($n = 65$). On obtient au final 388120 sondes qui passent le contrôle qualité (voir figure 3.8).

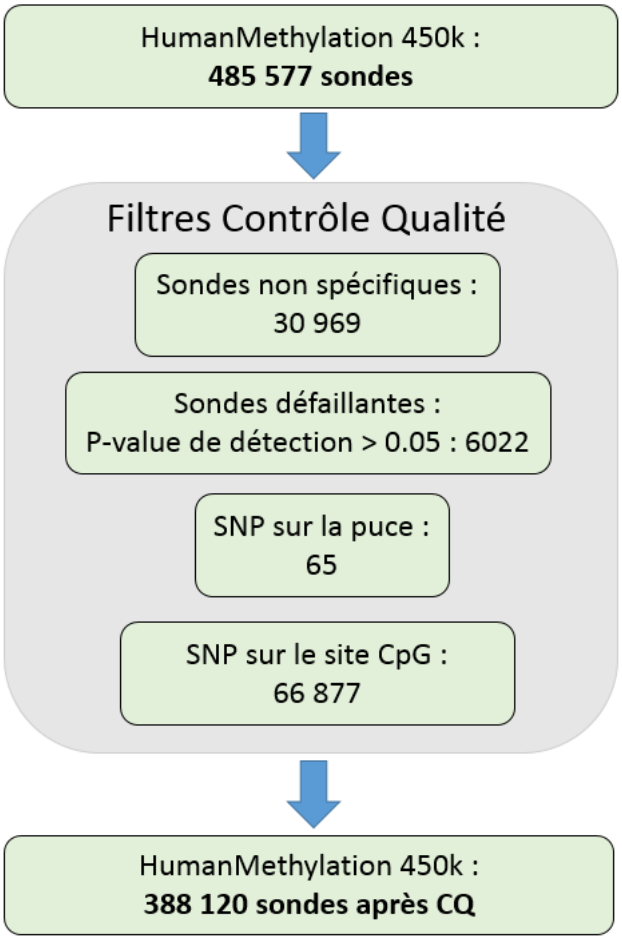

FIGURE 3.8 – Diagramme de flux du contrôle qualité pour les études MARTHA et F5L-Pedigrees.

Une fois l'exclusion des échantillons et des sondes problématiques, il est nécessaire de normaliser les données pour corriger les divers biais présents.

7 Les différents biais rencontrés

Divers biais peuvent être observés avec la technologie de puce à ADN. Certains communs à l'utilisation de puce à ADN (ex : le biais du bruit de fond), d'autres spécifiques à la puce HM450k (ex : l'utilisation de 2 types de sondes Infinium) et d'autres encore spécifiques aux conditions d'utilisation de la puce HM450k (la contamination cellulaire). Au cours de ce chapitre, nous allons parler de certains biais qui nécessitent de transformer les données pour les corriger comme le biais du bruit de fond (*Background noise* en anglais), celui lié à l'utilisation de 2 types de sondes (Infinium I vs Infinium II) ainsi que le biais du fluorochrome (*Dye bias*). Nous allons également parler d'autres biais qui ne seront pas corrigés comme le biais lié au taux de GC de la sonde (*GC content*) et d'autres qui seront

pris en compte dans les modèles statistiques comme le biais lié à l'effet de lot (*Batch effect* ou *Chip effect*) ou encore celui lié à la contamination cellulaire.

Bien qu'il soit nécessaire de corriger les divers biais présents pour minimiser les variations techniques, il est important de ne pas réaliser de correction trop sévère qui risquerait de diminuer également les variations biologiques et qui induirait un nouveau biais : l'excès de normalisation.

7.1 Biais du bruit de fond

Ce biais est dû à des hybridations non spécifiques aléatoires entre les sondes et les fragments d'ADN. Des fragments d'ADN n'ayant pas la séquence complémentaire de la sonde s'hybrident à celle-ci ce qui provoque une élongation et une fluorescence non désirée. Le bruit de fond peut être de deux sortes :

— un bruit sans structure augmentant globalement le niveau moyen des mesures et leur variabilité ;

— un bruit structuré, local, qui est dépendant de la position de la sonde sur la puce (également appelé biais spatial).

Sur les puces Illumina, la nature aléatoire du positionnement des billes rend les mesures robustes aux biais spatiaux. Toutefois une méthode de correction a été proposée pour corriger ce type de biais sur l'ancienne puce HM27k (Sabbah *et al.* 2011).

Le biais du bruit de fond peut être estimé de différentes façons.

Estimation par la méthode des contrôles négatifs

Ce biais est visible grâce à l'utilisation des 614 sondes de contrôles négatifs par échantillon. Ces sondes étant élaborées pour ne pas avoir de séquence complémentaire dans l'ADN et donc ne pouvant pas s'hybrider avec des fragments d'ADN. L'hybridation ne se faisant pas, aucun signal ne devrait être émis. Le signal détecté pour ces sondes correspond donc à la présence d'une hybridation non spécifique. Elles ont été conçues pour être équivalentes en termes de propriétés thermodynamiques aux sondes mesurant la méthylation.

Ce biais est visible sur la figure 3.9 qui représente la fonction de densité des intensités des sondes de contrôles négatifs.

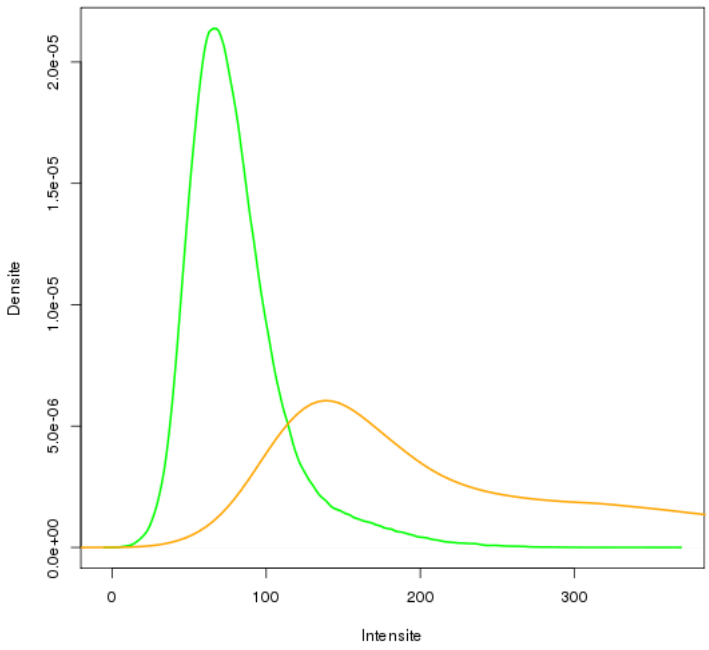

FIGURE 3.9 – Fonction de densité des intensités des sondes de contrôles négatifs obtenue à partir des 588 échantillons analysés. En vert, la fonction de densité des intensités des contrôles négatifs émises dans le vert (Cy3) et en jaune, la fonction de densité des intensités des contrôles négatifs émises dans le rouge (Cy5).

Estimation par la méthode de la valeur dominante des intensités

En posant l'hypothèse que la majorité des sites CpG sont non méthylés, la valeur dominante (le mode) des intensités obtenues par les sondes mesurant l'état méthylé correspond à la valeur du bruit de fond. L'estimation du bruit de fond par cette méthode repose sur l'utilisation des intensités obtenues par 135501 sondes Infinium I (46298 pour Cy5 et 89203 pour Cy3). Il s'agit de la méthode d'estimation du bruit de fond utilisée lorsque les sondes de contrôles négatifs ne sont pas disponibles.

Estimation par la méthode des intensités "out-of-band"

De par leur conception, les sondes de type Infinium I n'utilisent qu'un seul fluorochrome pour mesurer les deux états d'une cytosine. Toutefois le scanner d'Illumina mesure également la fluorescence du fluorochrome non utilisé. Ces intensités de fluorescence ne sont pas inutiles, car elles peuvent être utilisées pour estimer le bruit de fond et sont dites "out-of-band" (TRICHE *et al.* 2013).

L'estimation du bruit de fond par les intensités "out-of-band" est plus puissante et permet une estimation plus précise que celle se basant sur les intensités des sondes de contrôles négatifs. La méthode se basant sur les sondes de contrôles négatifs estime le bruit de fond via seulement 614

sondes or avec la méthode "out-of-band" l'estimation se base sur 178406 sondes ($2 * 89203$ sondes Infinium I mesurées avec Cy5) pour estimer le bruit de fond du fluorochrome Cy3 et 92596 sondes ($2 * 46298$ sondes Infinium I mesurées avec Cy3) pour le bruit de fond du fluorochrome Cy5. Nous pouvons voir sur la figure 3.10 que les histogrammes représentant le bruit de fond obtenus à partir des intensités "out-of-band" sont plus détaillés comparé aux histogrammes obtenus à partir des sondes de contrôles négatifs, cela reflète une estimation plus fine du bruit de fond.

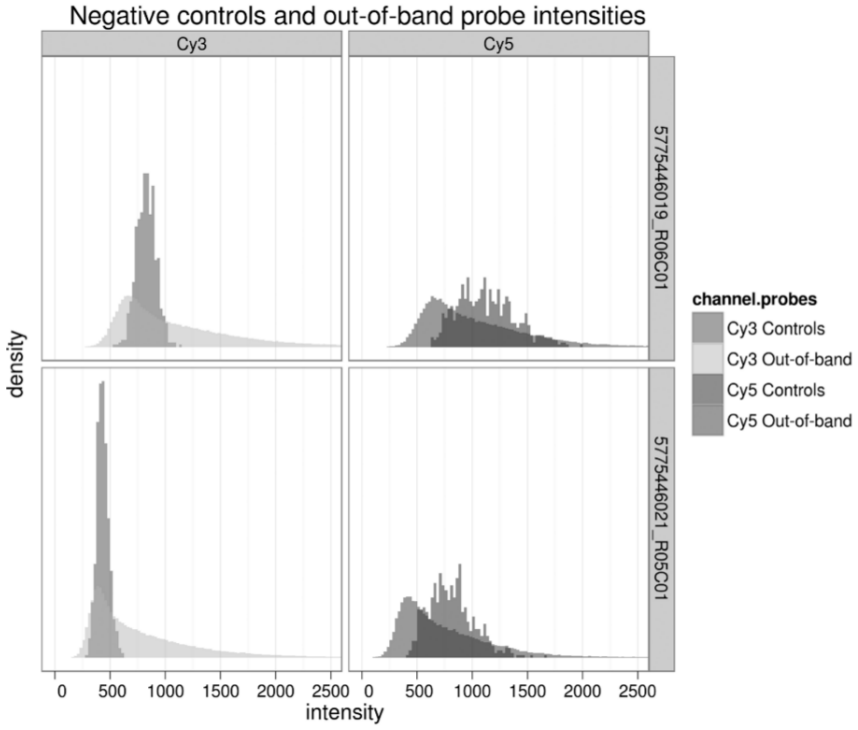

FIGURE 3.10 – Comparaison des estimations du bruit de fond réalisées par les sondes de contrôles négatifs et par les intensités "out-of-band". Figure provenant de TRICHE *et al.* 2013.

7.2 Biais lié à l'utilisation de 2 types de sondes

L'utilisation de deux technologies pour mesurer la méthylation de l'ADN au sein de la biopuce Illumina HumanMethylation450k engendre un biais. En effet, en comparant les résultats obtenus par la puce HM450k avec ceux obtenus par la méthode de référence (le pyroséquençage au bisulfite), les sondes de type Infinium II se révèlent être moins précises et moins reproductibles que les sondes de type Infinium I utilisées également dans la puce de génération précédente HM27k (DEDEURWAERDER *et al.* 2011). La figure 3.11, nous montre les densités des β en fonction du type de sonde Infinium I et Infinium II.

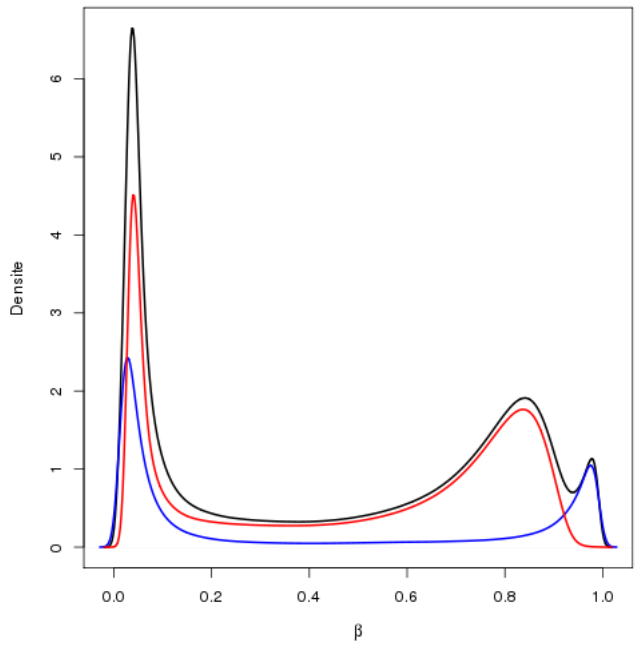

FIGURE 3.11 – Fonction de densité des β totaux (noire) et provenant des sondes de type I (bleue) ou II (rouge) pour les 588 échantillons.

La courbe noire représente la distribution de la totalité des β, la courbe bleue représente la distribution des β mesurés avec les sondes de type I et la courbe rouge celle des β mesurés avec les sondes de type II pour les 588 échantillons provenant des études MARTHA et F5L-Pedigrees. On peut voir sur la figure 3.11 que les pics à droite (correspondant à l'état méthylé) rouge (Infinium II) et bleu (Infinium I) ne sont pas superposables, ce qui illustre la différence de précision et reproductibilité en fonction du type des sondes.

Ce biais pose problème si l'on désire comparer un site mesuré avec Infinium 1 et un site mesuré avec Infinium 2. La puce HM450k n'ayant pas été conçue dans cette optique, Illumina ne propose pas de méthode de correction pour ce biais (voir la FAQ d'Illumina). Cette différence entre les deux types de sondes est également problématique lorsque l'on veut étudier la variabilité de la méthylation des sites CpG. On risque d'avoir un enrichissement de sites CpG variables mesurés avec les sondes Infinium II par rapport aux sites CpG mesurés avec les sondes Infinium I. Il en est de même si l'on s'intéresse aux sites CpG dont la méthylation est peu variable. Dans ce cas de figure, on risque d'avoir un enrichissement de sites CpG ayant été mesurés par des sondes Infinium I.

7.3 Biais du fluorochrome

La méthode de détection de la méthylation des sondes Infinium II engendre un biais car le fluorochrome utilisé pour détecter l'état méthylé (Cy3 : vert) diffère de celui utilisé pour mesurer l'état non méthylé (Cy5 : rouge). Ces deux fluorochromes étant chimiquement différents, ils ne possèdent pas la même efficacité d'hybridation. Le biais du fluorochrome est illustré sur la figure 3.12.

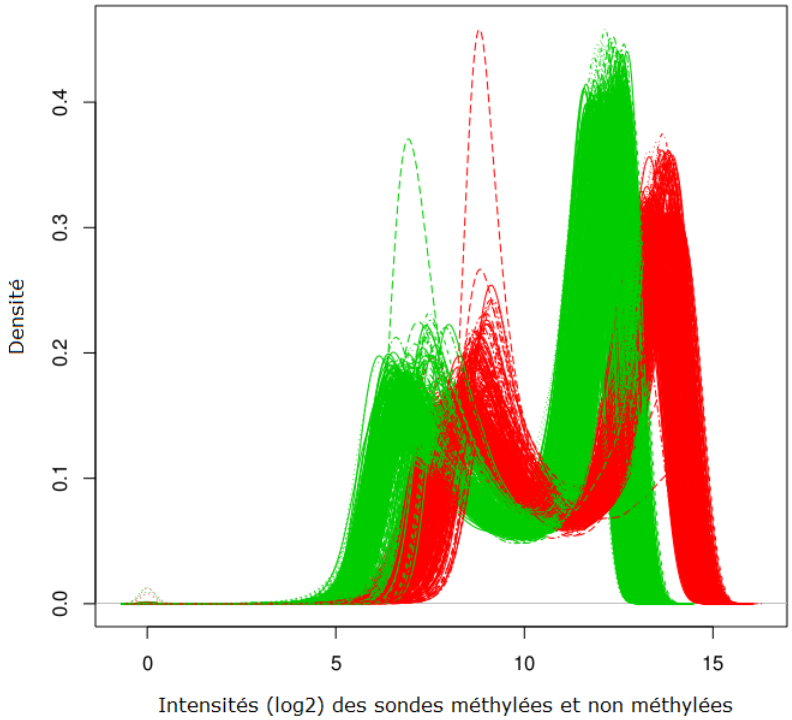

FIGURE 3.12 – Fonction de densité des 588 logarithmes binaires des intensités Infinium II pour les états méthylés (verts) et non méthylés (rouges).

On voit sur la figure 3.12 les fonctions de densités des logarithmes binaires des intensités obtenues pour les 588 échantillons. Les fonctions de densités des sondes pour les états méthylés (verts) sont décalés vers la gauche par rapport aux fonctions de densités des sondes pour les états non méthylés (rouges). Ce décalage résulte du biais du fluorochrome, car les intensités obtenues avec les deux fluorochromes devraient avoir la même fonction de répartition ce qui n'est pas le cas ici.

Ce biais est à l'origine, tout du moins en partie, du biais lié à l'utilisation de deux types de sondes. En effet, ce biais est systématiquement en faveur d'un des deux états mesurés par les sondes de type Infinium II qui utilisent un fluorochrome par état (Cy3 pour méthylé et Cy5 pour non méthylé). Alors que les sondes Infinium I utilisent aléatoirement le fluorochrome Cy3 ou le fluorochrome Cy5.

7.4 Biais du taux de GC dans la sonde

Les différentes sondes utilisées dans la puce HM450k possèdent un nombre différent de cytosine et de guanine. Il a été démontré que des sondes contenant un nombre élevé de cytosine et de guanine sous-estiment la valeur du β en raison de propriétés thermodynamiques différentes (Kuan *et al.* 2010). Ce biais serait plus important pour les sondes de type Infinium I car elles possèdent un taux de CG plus important que les sondes de types Infinium II (Maksimovic *et al.* 2012). La figure 3.13 est extraite de l'article de Kuan *et al.* 2010 et représente le biais "GC content", il s'agit de loci méthylés donc théoriquement avec une valeur β proche de 1. On voit sur le graphique ci-dessous que plus le taux de GC est important moins la valeur β est importante.

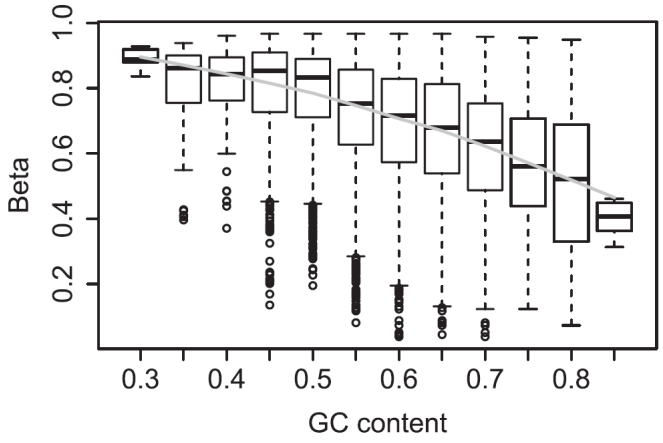

Figure 3.13 – Diagramme en boîte des valeurs β mesurées pour des loci méthylés en fonction du nombre de CG contenus dans la sonde. Graphique extrait de l'article de Kuan *et al.* 2010.

7.5 Biais lié à l'effet de lot

La capacité actuelle des lames (ou *chips* en anglais) Illumina utilisées permet de ne traiter simultanément que 12 sujets. Le scanner Illumina peut quant à lui mesurer simultanément l'intensité d'un maximum de 8 lames soit 96 échantillons. Les conditions expérimentales (manipulation/manipulateur) lors d'une mesure ne pouvant être reproduites à l'identique, il est important de prendre en compte la variabilité inter-puce pour éviter tout artefact, ce que l'on retrouve parfois sous la dénomination "*batch effect*". La mesure de la méthylation sur l'ensemble de nos 588 échantillons a nécessité l'utilisation d'un total de 50 lames. Lors des manipulations de ces 50 lames et lors de la mesure de l'intensité de fluorescence par le scanner, les conditions expérimentales ont pu varier. La variabilité entre lots est illustrée par la figure 3.14 qui fournit les diagrammes en boites des distributions des valeurs β non corrigées séparément par lots. La figure 3.14 nous montre que la variabilité des valeurs β diffère entre

les différentes puces.

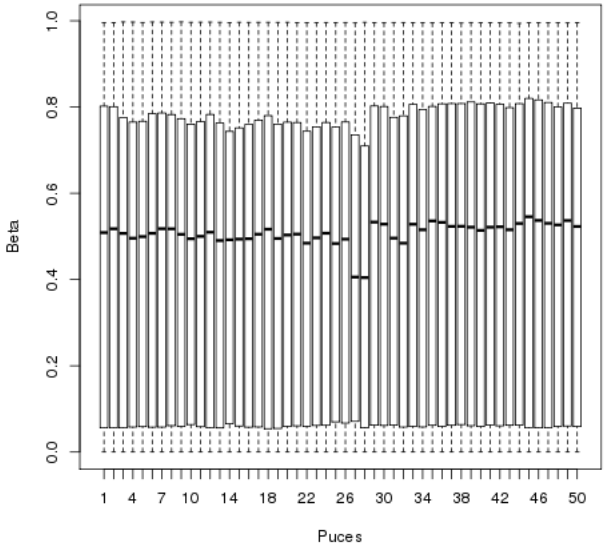

FIGURE 3.14 – Diagrammes en boites des valeurs β non corrigées par puces (ou lots) pour les 588 échantillons.

Des variations similaires à moindre effet peuvent également s'appliquer au niveau des lames, on parlera alors de l'effet *chip*. Les lames peuvent contenir 12 échantillons répartis en deux colonnes de 6 échantillons. Chaque position sur la micropuce peut avoir une variabilité différente non désirée provoqué par exemple par des caractéristiques du scanner utilisé pour mesurer la fluorescence.

7.6 Biais de la contamination cellulaire

Il s'agit d'un biais spécifique à l'analyse d'échantillons contenant un mélange de type cellulaire comme c'est le cas dans ce travail de thèse où la méthylation de l'ADN a été mesurée sur du sang périphérique.

La mesure de la méthylation de l'ADN dans le sang se fait sur les cellules sanguines possédant un noyau dont les leucocytes. Comme expliqué dans les parties précédentes, l'extraction de l'ADN est réalisée à partir des cellules circulantes du sang périphérique, les cellules circulantes étant composées d'érythrocytes, de leucocytes et de thrombocytes. Or les seules à posséder un noyau et donc de l'ADN sont les leucocytes, elles-mêmes composées de différents types cellulaires (neutrophiles, éosinophile, basophile, lymphocyte et monocyte). Il a été démontré que les différents leucocytes possédaient des profils de méthylations différents (Y. V. SUN *et al.* 2010), donc des proportions différentes entre les échantillons de ces différents types cellulaires conduisent à mesurer des profils de méthylation différents. Il est donc important de prendre en compte les proportions des leucocytes dans le sang lors

des analyses (Jaffe & Irizarry 2014 ; Y. Liu *et al.* 2013 ; Michels *et al.* 2013 ; Paul & S. Beck 2014). Cela peut se faire par correction ou ajustement lors des analyses.

Lorsque les proportions des différents types cellulaires n'ont pas été mesurées, il est toutefois possible de les estimer via une méthode de déconvolution (Houseman, Accomando *et al.* 2012 ; Jaffe & Irizarry 2014 ; Koestler, Christensen *et al.* 2013) à partir d'une base de données de référence contenant les données de méthylation mesurée sur des échantillons purs des différents types cellulaires. Les niveaux de méthylation d'échantillons purs ont été mesurés pour les cellules sanguines (Reinius *et al.* 2012) et pour les cellules du cortex préfrontal (Guintivano *et al.* 2013). À partir d'échantillons purs, c'est-à-dire ne contenant qu'un unique type cellulaire, des profils de méthylation différentielles ont été établis. Entre 100 et 500 sites CpG ont été identifiés pour avoir un profil de méthylation distinct pour chaque type cellulaire. Pour chaque site identifié, la proportion mesurée de l'ADN méthylé est supposée augmenter comme un mélange linéaire des profils spécifiques des cellules distinctes. Les coefficients de mélange sont supposés représenter les proportions de chaque type cellulaire. Pour plus de détails sur la méthode statistique, j'invite le lecteur à se référer à l'article de Houseman, Accomando *et al.* 2012. D'autres méthodes existent pour corriger les données de méthylation en prenant en compte les différentes proportions cellulaires (Gagnon-Bartsch & Speed 2012) ou pour estimer les proportions cellulaires sans les niveaux de méthylation de référence des différents types cellulaires présent dans l'échantillon étudié (Houseman, Kelsey *et al.* 2015 ; Houseman, Molitor *et al.* 2014 ; Zou *et al.* 2014).

Dans l'étude MARTHA, les quantités leucocytaires ont été mesurées via l'analyseur d'hématologie cellulaire ADVIA® 120 Hematology System (Siemens Healthcare Diagnostics, Deerfield, IL). Nous disposons des quantités mesurées des types cellulaires suivant : lymphocyte, monocyte, basophile, éosinophile et neutrophile.

Alors que dans l'étude F5L-Pedigrees, les proportions leucocytaires ont été estimées via la méthode de déconvolution proposé par Houseman, Accomando *et al.* 2012 en utilisant comme base de référence celle de Reinius *et al.* 2012. Pour évaluer l'efficacité de cette méthode, j'ai comparé les proportions mesurées dans l'étude MARTHA aux proportions estimées dans cette même étude par l'algorithme de Houseman. Les figures 3.15, 3.16 et 3.17 représentent les comparaisons entre les proportions mesurées et les proportions estimées respectivement pour les granulocytes, les lymphocytes et les monocytes.

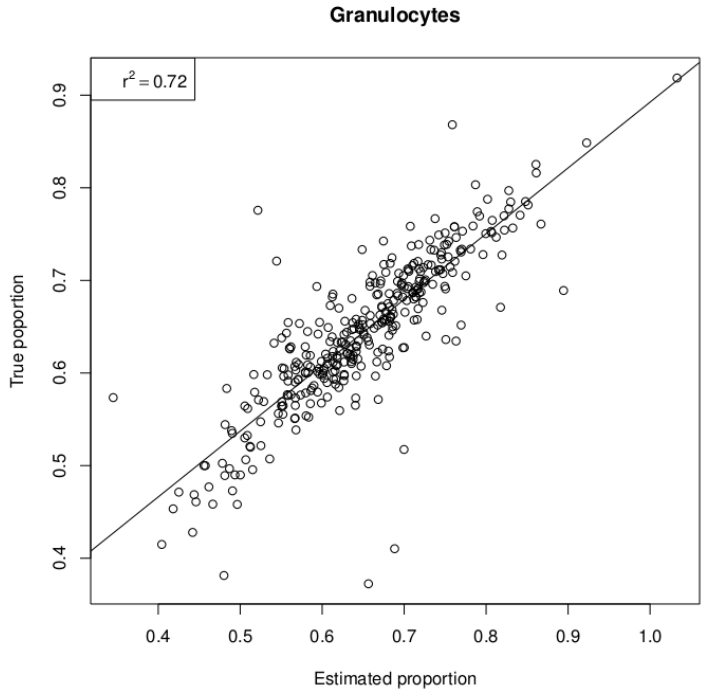

FIGURE 3.15 – Comparaison des proportions mesurées de granulocyte dans l'étude MARTHA aux proportions estimées dans cette même étude par l'algorithme de Houseman.

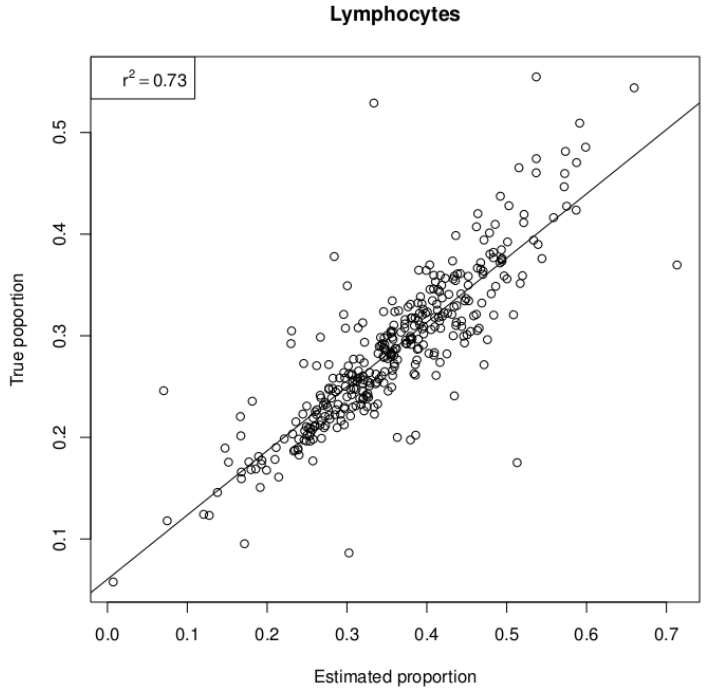

FIGURE 3.16 – Comparaison des proportions mesurées de lymphocyte dans l'étude MARTHA aux proportions estimées dans cette même étude par l'algorithme de Houseman.

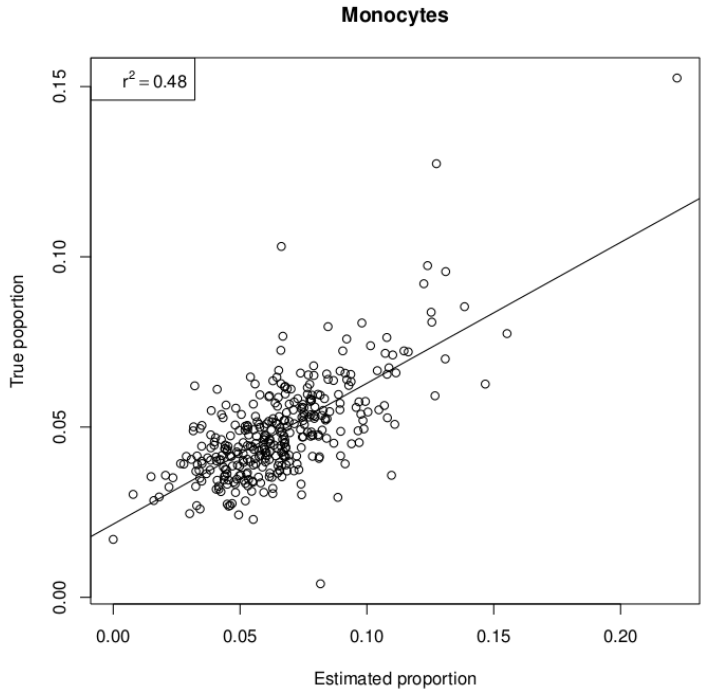

FIGURE 3.17 – Comparaison des proportions mesurées de monocyte dans l'étude MARTHA aux proportions estimées dans cette même étude par l'algorithme de Houseman.

Nous obtenons une bonne corrélation entre les valeurs mesurées et les valeurs estimées. La corrélation est d'environ 0,7 pour les granulocytes et les lymphocytes et d'environ 0,5 pour les monocytes. Ces résultats nous permettent de valider que l'algorithme de Houseman fonctionne bien et que les valeurs estimées dans l'étude F5L-Pedigrees sont fiables. Nous disposons pour l'étude F5L-Pedigrees, des quantités estimées des types cellulaires suivant : lymphocyte T CD4+, lymphocyte T CD8+, lymphocyte NKT, lymphocyte B, monocyte et granulocyte.

Les proportions de ces différents types cellulaires sont ajoutées lors des analyses comme des variables d'ajustements (voir la partie sur les analyses statistiques des données de méthylation, p. 92).

8 Méthodes de Correction & Normalisation

Chacun des biais mentionnés nécessite des méthodes particulières pour les corriger. Je présente ici les principales méthodes couramment utilisées ainsi que quelques améliorations que j'ai essayé d'apporter. Certaines méthodes sont dites "interpuces" et d'autres sont dites "intrapuces". Le type de la méthode (interpuce ou intrapuce) est fortement lié au biais qu'il corrige. En effet pour corriger l'effet de lot, il est nécessaire de réaliser une normalisation "interpuce" pour ajuster les intensités

obtenues avec les différentes puces. Alors que pour corriger le bruit de fond qui lui peut varier entre les différentes puces, une méthode "intrapuce" est plus adaptée.

8.1 Correction du bruit de fond

Le premier biais étudié est celui appelé "bruit de fond". Il peut être corrigé simplement en soustrayant la valeur de l'intensité du bruit de fond aux intensités des sondes de mesures (méthode dite "standard") ou en modélisant le signal observé comme résultant du signal réel et du bruit de fond (méthode dite "NormExp").

Méthode standard de soustraction

Parmi les différentes méthodes existantes pour corriger le bruit de fond, la méthode la plus simple est de quantifier le bruit de fond par échantillon et par canal (Cy3 et Cy5) puis de soustraire la valeur estimée du bruit de fond (par exemple à partir des sondes de contrôles négatifs, voir la partie sur l'estimation du bruit de fond 7.1) aux valeurs des intensités des sondes mesurant la méthylation. La méthode de correction proposé par Illumina dans le logiciel GenomeStudio se base sur ce principe. Elle estime le bruit de fond par la méthode des sondes de contrôles négatifs. À partir des intensités des sondes de contrôles négatifs, elle en déduit une valeur de référence du bruit de fond qui correspond au $5^{\text{ème}}$ centile des intensités des sondes de contrôles négatifs. On déduit ensuite la valeur de référence du bruit de fond aux valeurs des intensités des sondes mesurant la méthylation. En cas d'apparition de valeurs négatives, celles-ci sont remplacées par 0. Une variante de cette dernière étape est de remplacer les valeurs négatives non pas par 0 mais par la valeur de référence du bruit de fond.

Différentes alternatives pour déterminer la valeur de référence du bruit de fond ont également été testées :

- $5^{\text{ème}}$ centile : méthode proposée par Illumina dans GenomeStudio
- $20^{\text{ème}}$ centile
- médiane
- moyenne
- mode : méthode proposée par le package "lumi", (voir la partie sur l'estimation du bruit de fond par la méthode du mode, p. 47)

La correction du bruit de fond se fait pour chaque individu indépendamment ainsi que pour chaque canal (Cy3 et Cy5) de mesure de la fluorescence indépendamment. C'est-à-dire que pour chaque

individu, il est nécessaire d'appliquer deux corrections indépendantes pour chaque type d'intensités mesurées (intensités vertes et intensités rouges).

Illumina recommande d'analyser les données avec ou sans correction du bruit de fond en fonction des résultats obtenus. La correction du bruit de fond est nécessaire lorsque l'on compare des données provenant des différents types de scanners car il peut y avoir des différences techniques : par exemple, l'iScan et HiScan ont des correcteurs très différents. La correction du bruit de fond à moins d'effet et peut être inutile lorsque les puces sont analysées sur le même scanner.

La figure 3.18 représente les intensités des différents signaux obtenus après ces différentes corrections.

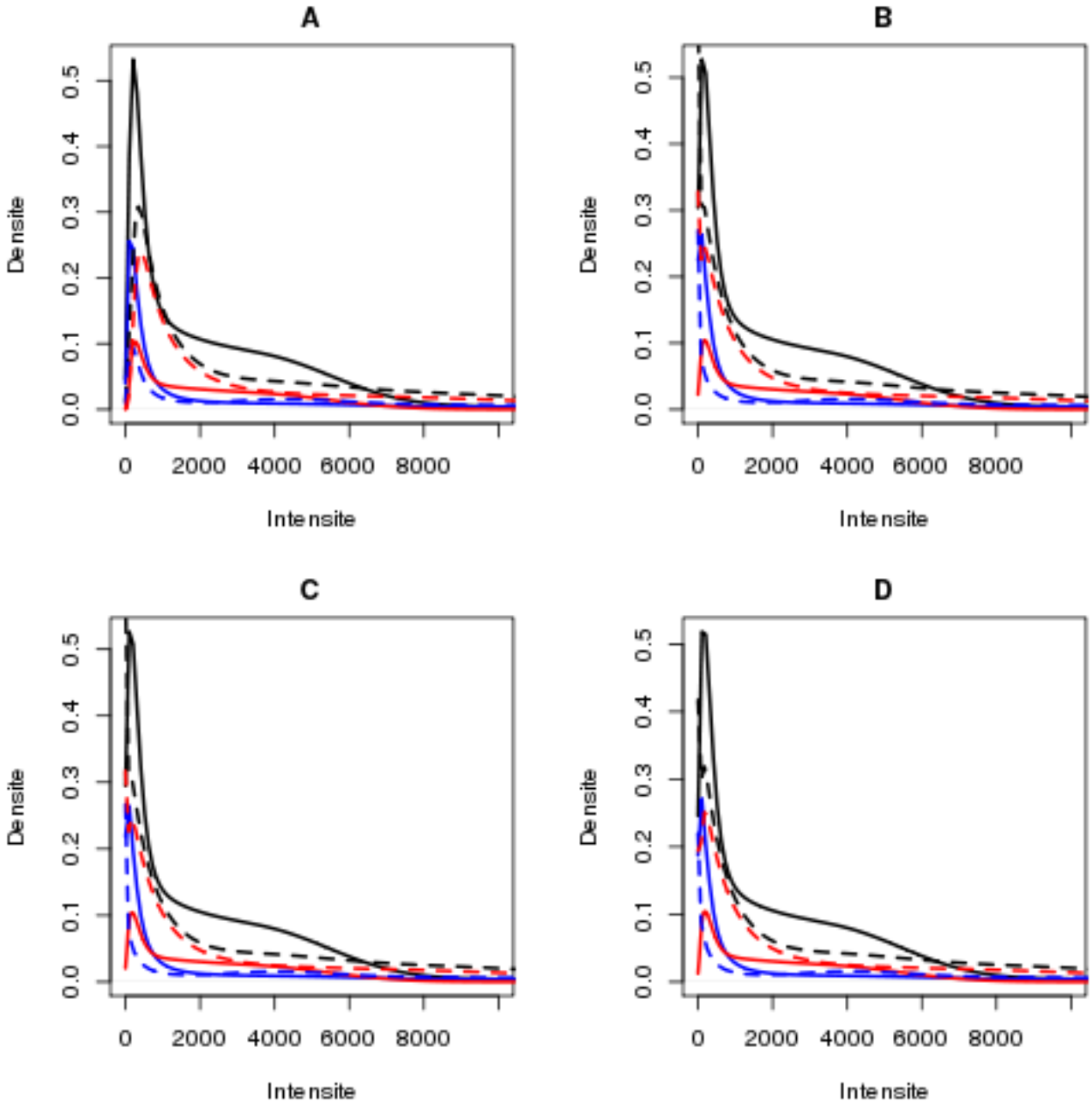

FIGURE 3.18 – Fonction de densité des intensités des différents signaux obtenus à partir des 588 échantillons analysés. Le graphique A correspond aux données brutes, le B correspond aux données traitées par la méthode de la moyenne, le C correspond aux données traitées par la méthode la médiane et le D correspond aux données traitées par la méthode du 20$^{\text{ème}}$ centile. En continue, les courbes des fonctions de densités des états méthylés et en pointillé, les courbes des fonctions de densités des états non méthylés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

Les fonctions de densités des intensités sont modifiées par la correction du bruit de fond. La soustraction des intensités par la valeur de référence du bruit de fond va diminuer les intensités d'où

une légère translation horizontale vers la gauche des différentes fonctions de densités des figures 3.18 B, C et D. La translation est moins prononcée pour la correction par la méthode du $20^{\text{ème}}$ centile car on soustrait les intensités par une valeur plus faible. La correction des intensités va se répercuter sur les valeurs des β, la figure 3.19 représentant les fonctions de densités des β après les diverses corrections.

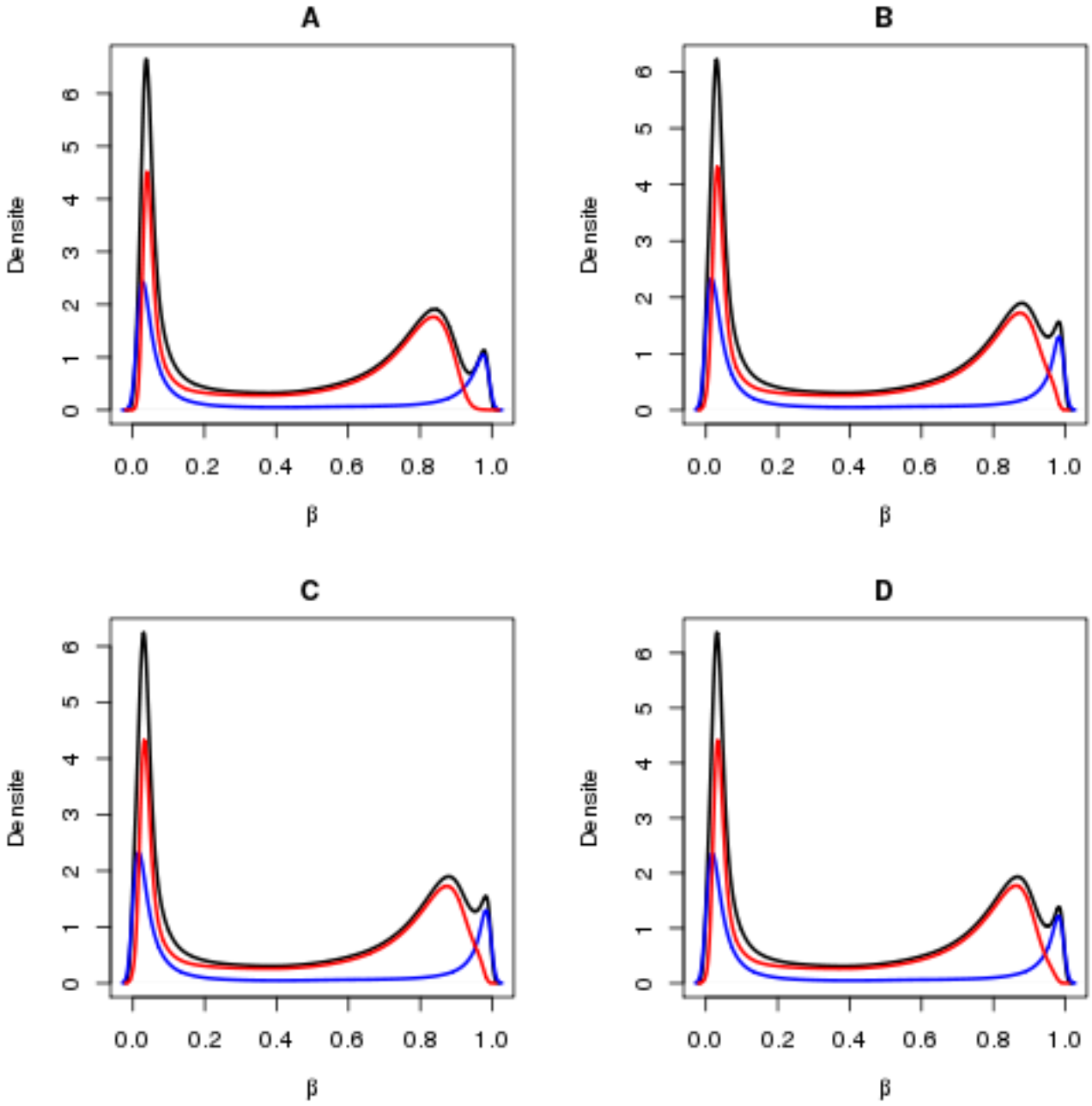

FIGURE 3.19 – Fonctions de densité des β obtenue à partir des 588 échantillons analysés. Le graphique A correspond aux données brutes, B correspond aux données traitées par la méthode de la moyenne, le C correspond aux données traitées par la méthode la médiane et le D correspond aux données traitées par la méthode du $20^{\text{ème}}$ centile. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

Il n'y a pas de grosses différences sur les fonctions de densité des β avant et après correction du bruit de fond par les différentes méthodes testées. On note toutefois que les pics des courbes des densités des β provenant des sondes de type II semblent avoir une pente plus douce au environ de

$\beta = 1$.

Méthode de modélisation "NormExp"

Une méthode proposée par Irizarry, Hobbs *et al.* 2003 consiste à modéliser les intensités observées X comme la résultante d'un signal S distribué selon une loi exponentielle de moyenne α et d'un bruit de fond B distribué selon une loi normale de moyenne μ et de variance σ^2. Estimer le vrai signal pour une sonde donnée revient à estimer l'espérance de S sachant l'intensité observée X donnée par la formule suivante :

$$\mathbb{E}(S|X = x) = \mu_{S.X} + \frac{\sigma^2 \phi(0; \mu_{S.X}, \sigma^2)}{1 - \Phi(0; \mu_{S.X}, \sigma^2)}$$

où ϕ et Φ désignent respectivement la densité et la fonction de répartition d'une loi normale et avec $\mu_{S.X} = x - \mu - \frac{\sigma^2}{\alpha}$.

σ et μ pouvant être estimés directement à partir des contrôles négatifs et α par maximum de vraisemblance, on déduit facilement l'espérance du vrai signal (positif par construction) à partir des observations. Cette normalisation doit être réalisée indépendamment pour les signaux des deux fluorochromes (Cy3 et Cy5).

La variante de la méthode "NormExp" utilisée sur nos données consiste en l'ajout d'une compensation (ou "offset") de 50 à l'intensité estimée des sondes (Ritchie *et al.* 2007). Il s'agit d'une méthode simple de stabilisation de la variance analogue à l'approche décrite par Rocke & Durbin 2003 qui permet d'obtenir des intensités éloignées de 0.

$$X_{corrigee} = X_{estimee} + 50$$

Où $X_{estimee}$ est la valeur de l'intensité estimée via la méthode "NormExp", 50 la valeur de la compensation et $X_{corrigee}$ la valeur de l'intensité corrigée.

Une alternative est d'utiliser la méthode "NormExp" en estimant le bruit de fond non pas avec les intensités des sondes de contrôles négatifs mais avec les intensités dites "out-of-band" (voir la partie sur l'estimation du bruit de fond par la méthode "out-of-band", p. 47), méthode proposée dans le package "methylumi" et appelée "Noob". L'estimation des paramètres μ (moyenne) et σ^2 (variance) de la loi normale représentant la composante du bruit de fond de l'intensité mesurée ne se fera pas à partir des intensités obtenues via les 614 sondes de contrôles négatifs mais à partir des intensités "out-of-band" de 271002 sondes.

Autres méthodes de correction

Il existe d'autres méthodes pour corriger le bruit de fond non testées dans ce travail de thèse. Les efficacités de certaines d'entre elles sont comparées dans l'article de Ritchie *et al.* 2007.

8.2 Correction du biais lié à l'utilisation de deux types de sondes

Le second biais qui a été étudié est le biais lié aux deux types de sondes utilisées (Infinium I & Infinium II).

Méthode "Peak-based correction"

La première méthode "Peak-based correction" (ou PBC) (Dedeurwaerder *et al.* 2011) s'applique en 4 étapes.

La première étape consiste à convertir les valeurs β en valeurs M via la fonction logit.

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

La figure 3.20 montre les fonctions de densités des β à gauche (avant la première étape) et des valeurs M à droite (après la première étape).

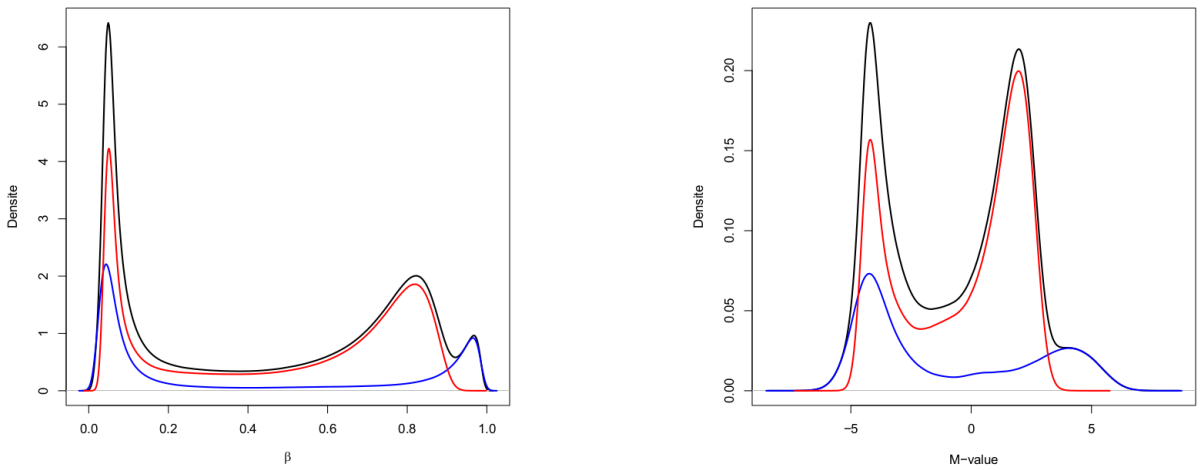

Figure 3.20 – Fonctions de densité des β (gauche) et des valeurs M (droite) obtenues à partir des 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

On voit sur la figure 3.20 que les valeurs M sont centrées en 0.

On applique ensuite une deuxième étape qui consiste à aligner les pics des deux courbes en -1 pour les pics de non méthylation et en 1 pour les pics de méthylation comme illustré sur la figure 3.21 ci-dessous. Pour cela, on va diviser les valeurs M négatives de Infinium I par la distance du pic de non méthylation de Infinium I à 0 ce qui va positionner le pic en -1. On procède de la même façon avec les valeurs M négatives de Infinium II qui seront divisées par la distance du pic de non méthylation de Infinium II à 0 ce qui va également positionner le pic en -1. On aura donc les deux pics de non méthylation (le pic Infinium I et le pic Infinium II) qui seront alignés en -1. On réalise la même méthode pour les pics de méthylation, c'est-à-dire qu'on va diviser les valeurs M positives de Infinium I par la distance du pic de méthylation de Infinium I à 0 ce qui va positionner le pic en 1. Puis on divise les valeurs M positives de Infinium II par la distance du pic de méthylation de Infinium II à 0 ce qui va également positionner le pic en 1. On aura ainsi les deux pics de méthylation (le pic Infinium I et le pic Infinium II) qui seront alignés en 1.

La figure 3.21 montre les fonctions de densités des valeurs M avant alignement des pics (à gauche) et après alignement des pics (à droite).

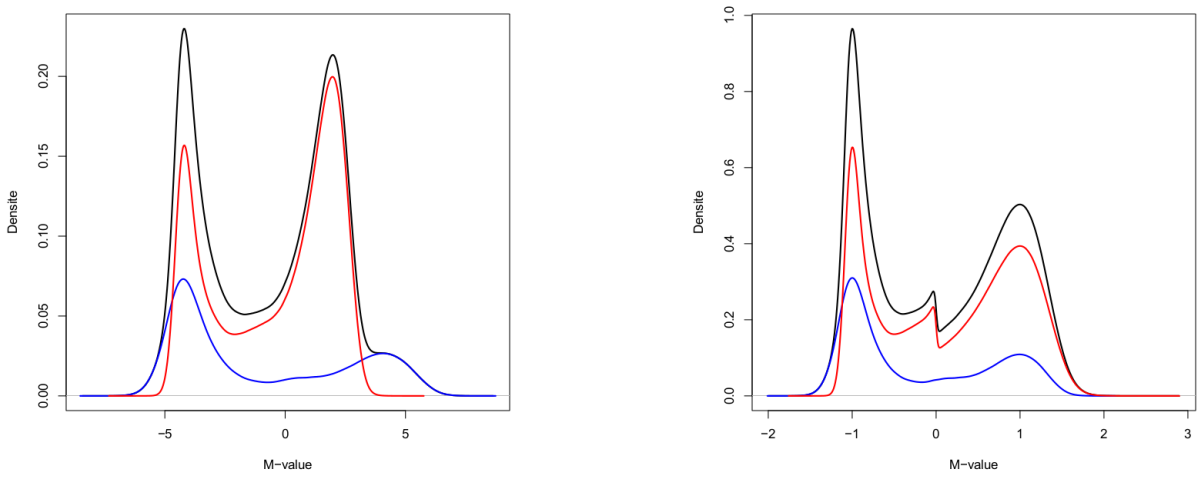

FIGURE 3.21 – Fonctions de densité des valeurs M brutes (gauche) et M ajustées (droite) pour les 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

La troisième étape consiste à repositionner les pics ainsi superposés à l'endroit où se situaient les pics des fonctions de densités des valeurs M provenant des sondes Infinium I car elles sont considérées comme plus fiables que les valeurs des sondes Infinium II et servent donc de référence pour l'alignement. Pour repositionner les pics, il suffit de multiplier les valeurs M négatives (de Infinium I et II)

par la distance qui sépare le pic de non méthylation de Infinium I et les valeurs M positives (de Infinium I et II) par la distance qui sépare le pic de méthylation de Infinium I. La figure 3.22 montre les fonctions de densités avant le repositionnement des pics (à gauche) et après le repositionnement des pics (à droite). En effet, on voit sur la figure de gauche que les pics sont centrés en -1 (non méthylés) et 1 (méthylés) alors que sur la figure de droite, ils sont centrés aux alentours de -5 (non méthylés) et de 5 (méthylés).

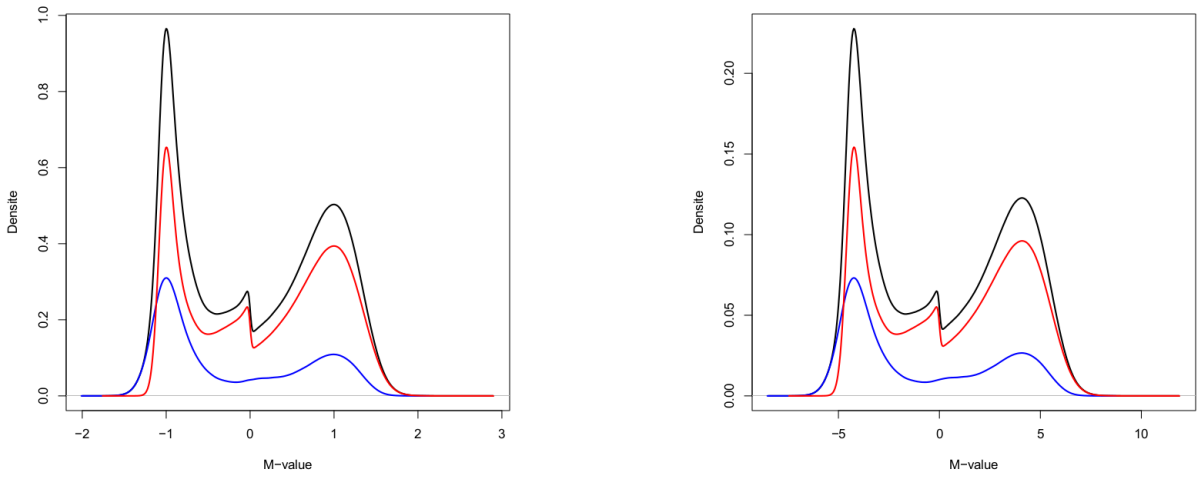

FIGURE 3.22 – Fonctions de densité des valeurs M ajustées (gauche) et M corrigées (droite) pour les 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

La dernière étape consiste à reconvertir les valeurs M corrigées en valeurs β corrigées via la formule :

$$\beta_{corr} = (\frac{2^{M_{corr}}}{2^{M_{corr}}+1})$$

La figure 3.23 montre les valeurs β corrigées par la méthode "PBC".

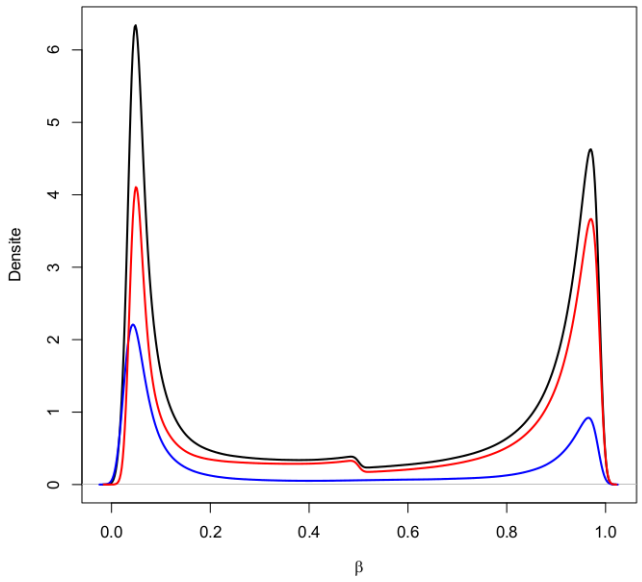

FIGURE 3.23 – Fonction de densité des valeurs β corrigées obtenues par la méthode "Peak-based correction" pour les 588 échantillons. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

On observe sur la figure 3.23 que le biais du type de sonde est corrigé mais on remarque également la présence d'une discontinuité en 0,5 sur la distribution des β provenant des sondes de types II. La présence de cette discontinuité est problématique car cela indique qu'il y aurait une plus forte probabilité d'obtenir un β inférieur à 0,5 et donc à un état non méthylé. Il a donc été entrepris d'améliorer cette méthode pour faire disparaître cette discontinuité. C'est dans ce but qu'a été réalisée la méthode "dnPeak-based correction".

Méthode "dnPeak-based correction"

Cette amélioration a été mise au point avec l'aide de Nicolas Greliche, un ancien doctorant du laboratoire. Elle est basée sur la méthode "PBC" (*Peak-based correction*) et est donc composée des 4 mêmes étapes, c'est-à-dire l'étape de transformation des β en M puis l'étape d'alignement des pics en -1 et 1, vient ensuite l'étape de repositionnement des pics en fonction de ceux obtenus par les sondes Infinium I et enfin la transformation des valeurs M corrigées en valeurs β corrigées. La méthode "dnPeak-based correction" diffère de la méthode "PBC" au niveau de l'étape d'alignement des pics en -1 et 1.

L'origine de la discontinuité présente dans la méthode "PBC" pourrait être liée à une trop forte variation de la correction appliquée aux données lors de l'étape de centrage des pics. On remarque sur la figure 3.24 ci-dessous qui représente les fonctions de densité des valeurs M ajustées (pics alignés) avec la méthode "PBC" l'apparition d'un pic sur la courbe de distribution des sondes de type II alors qu'il n'est pas présent sur la courbe des sondes de type I.

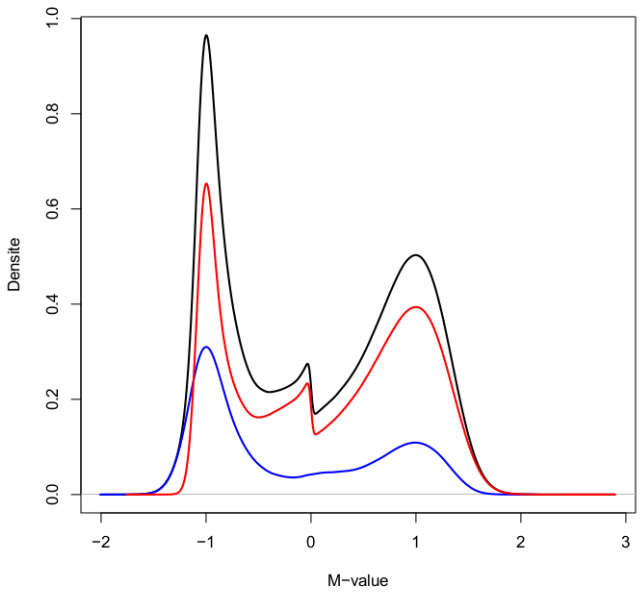

FIGURE 3.24 – Fonctions de densité des valeurs M ajustées avec la méthode "PBC" pour les 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

Cela est dû au changement brutal de la correction appliquée entre les valeurs positives et les valeurs négatives. Les valeurs inférieures à 0 vont subir un certain traitement tandis que les valeurs supérieures à 0 vont subir un traitement différent, ce qui va induire une discontinuité pour les valeurs aux alentours de 0. Nous pouvons voir sur la figure 3.25 qui illustre la fonction représentant la correction (dénominateur) utilisée dans la méthode "PBC" en fonction de la valeur M, le changement brutal de correction lorsque M change de signe.

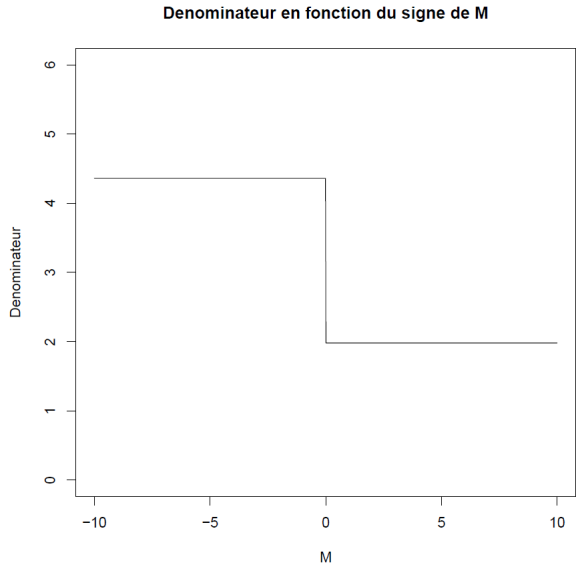

FIGURE 3.25 – Fonction représentant le changement de correction (dénominateur) utilisé dans la méthode "Peak-based correction" en fonction de la valeur M.

Nicolas Greliche a mis au point la fonction suivante dans le but d'avoir un changement moins brutal :

$$(arctan(k*x) - arctan(k*d_{NM})) * (\frac{d_{NM}+d_M}{arctan(k*d_M)-arctan(k*d_{NM})}) + |d_{NM}|$$

— k : un paramètre réglable, permettant d'accentuer ou d'adoucir la pente de la courbe.

— d_{NM} : la distance du pic de non méthylation à 0.

— d_M : la distance du pic de méthylation à 0.

La partie en bleue de la formule ci-dessus sert de base pour déterminer la correction à apporter. La partie en vert permet de positionner la courbe à la bonne ordonnée. La partie rose, quant à elle, permet d'avoir la bonne amplitude entre le minimum et le maximum de la courbe (elle sert à définir la distance entre les 2 pics).

Cette fonction permet d'avoir une correction proche de celle de "PBC" lorsque les valeurs de M sont éloignées de 0 et une correction qui s'en éloigne mais avec un changement moins brutal aux alentours de 0. La figure 3.26 illustre la fonction représentant le changement de correction (dénominateur) obtenu via la formule ci-dessus et utilisé dans la méthode "dnPeak-based correction" (dnPBC) en fonction de la valeur M.

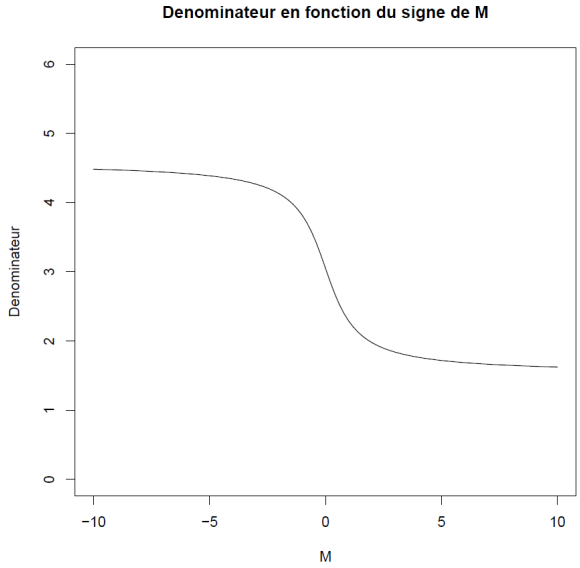

FIGURE 3.26 – Fonction représentant le changement de correction (dénominateur) utilisé dans la méthode "dnPeak-based correction" en fonction de la valeur M.

Une fois les pics alignés avec la méthode "dnPBC", il suffit d'appliquer les mêmes étapes que celles de la méthode "PBC" pour obtenir les valeurs β corrigées. La figure 3.27 montre les fonctions de densité des valeurs M ajustées avec "PBC" et avec "dnPBC".

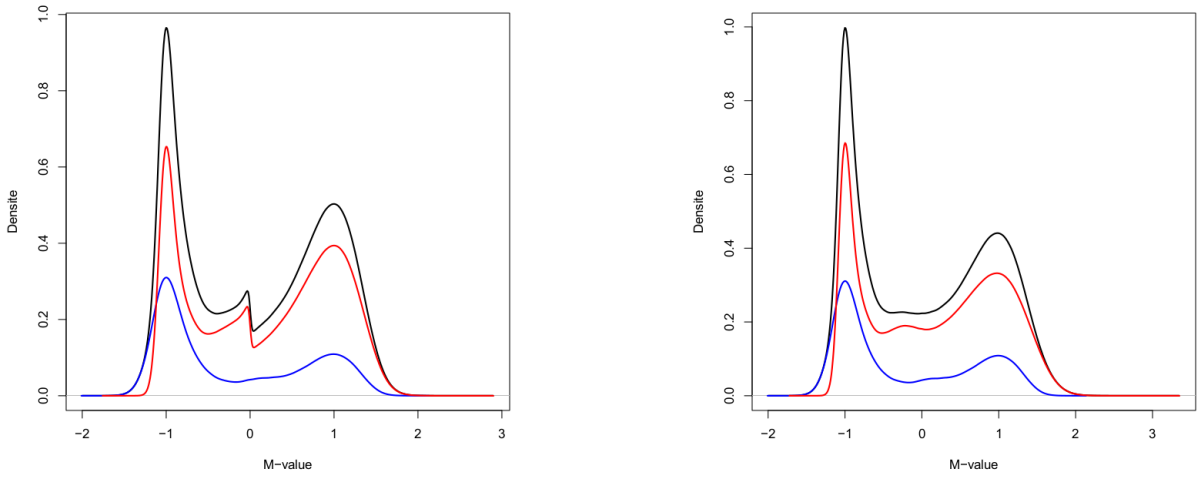

FIGURE 3.27 – Fonctions de densité des valeurs M ajustées avec "PBC" (gauche) et M ajustées avec "dnPBC" (droite) pour les 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

Nous pouvons voir sur cette figure que la correction "dnPBC" a fait disparaître la discontinuité.

La figure 3.28 montre les fonctions de densité des valeurs β corrigées avec "PBC" et avec "dnPBC".

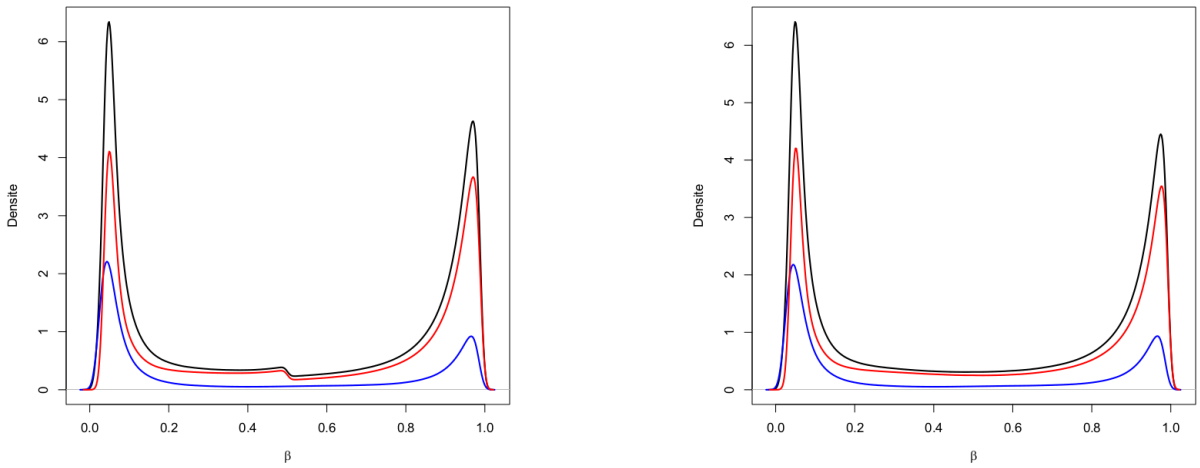

FIGURE 3.28 – Fonctions de densité des valeurs β corrigées avec "PBC" (gauche) et avec "dnPBC" (droite) pour les 588 échantillons analysés. La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

Nous pouvons voir sur ce graphique que la méthode "dnPBC" corrige bien le biais lié à l'utilisation de deux types de sonde et que, de plus, elle ne génère pas de discontinuité comme la méthode "PBC" pour les valeurs β proche de 0,5. La méthode "dnPBC" semble ici plus efficace que la méthode "PBC" pour corriger le biais lié à l'utilisation de deux types de sonde.

Méthode "SWAN"

Une autre méthode de correction du biais généré par l'utilisation de deux types de sondes a également été testée. La méthode "SWAN" pour "Subset Quantile Within-Array Normalization" (MAKSIMOVIC *et al.* 2012) est disponible dans le package R (Bioconductor) "minfi" (ARYEE *et al.* 2014). Le principe de "SWAN" est d'utiliser une normalisation par quantile entre les intensités de type Infinium I et de type Infinium II. Cette normalisation est réalisée pour les intensités A (non méthylée) et B (méthylée) séparément. La correction "SWAN" est à réaliser pour chaque individu indépendamment et se déroule en deux étapes.

La première étape consiste à déterminer une distribution moyenne des quantiles en utilisant des sous-groupes de sondes similaires sur leur contenu en CpG. Cette étape peut également être décomposée ainsi :

— Sélection aléatoire de N sondes Infinium I et II ayant 1, 2 et 3 site(s) CpG. N correspond au nombre de sondes dans le plus petit des 6 groupes ainsi créé. Si aucun filtre n'est appliqué

alors le nombre de sondes du plus petit groupe (groupe Infinium I ayant un site CpG) sera de 11 303. Chacun des 6 groupes ci-dessous a donc un nombre N de sondes.

- Infinium I ayant un site CpG
- Infinium I ayant deux sites CpG
- Infinium I ayant trois sites CpG
- Infinium II ayant un site CpG
- Infinium II ayant deux sites CpG
- Infinium II ayant trois sites CpG

— Pour chaque type de sonde, les intensités sont triées par ordre croissant. Chacun des 2 groupes ci-dessous a donc un nombre 3N de sondes.

- Infinium I : ayant un, deux et trois sites CpG trié par ordre croissant.
- Infinium II : ayant un, deux et trois sites CpG trié par ordre croissant.

— La moyenne entre la valeur de l'intensité de rang i (i peut prendre une valeur de 1 à 3N) de Infinium I et la valeur de l'intensité de rang i de Infinium II est attribuée comme valeur commune pour les intensité de rang i pour les 2 types de sonde.

Ces nouvelles valeurs d'intensités normalisées vont servir de référence pour déterminer les intensités des autres sondes qui n'ont pas été incluses à la précédente étape.

A partir des intensités normalisées à l'étape précédente, on détermine par interpolation linéaire les intensités des sondes restantes (majoritairement des sondes de type II) :

— On utilise la formule suivante pour déterminer les nouvelles valeurs :

$$f(x) = y_a + (x - x_a)\frac{y_b - y_a}{x_b - x_a}$$

- **x** : correspond à la valeur de l'intensité à normaliser.

  Pour normaliser cette valeur, il faut déterminer le rang de l'intensité mesurée de la sonde à normaliser en comparant son intensité avec les intensités normalisées précédemment. Une fois son rang établi, on utilise les intensités normalisées des 2 sondes dont le rang encadre la sonde à normaliser dans la formule ci-dessus.

- x_a : correspond à la valeur de l'intensité non normalisée de la sonde a de rang inférieur.
- y_a : correspond à la valeur de l'intensité normalisée de la sonde a de rang inférieur.
- x_b : correspond à la valeur de l'intensité non normalisée de la sonde b de rang supérieur.
- y_b : correspond à la valeur de l'intensité normalisée de la sonde b de rang supérieur.

— Dans le cas où il manque une sonde encadrant la sonde à normaliser, l'interpolation linéaire

est impossible. Il peut y avoir deux cas de figure :

— soit la valeur de l'intensité de la sonde à normaliser est supérieure à la valeur maximale des intensités mesurées des sondes de l'échantillon normalisé.

— soit la valeur de l'intensité de la sonde à normaliser est inférieure à la valeur minimale des intensités mesurées des sondes de l'échantillon normalisé.

Dans le premier cas, on calcule la différence entre l'intensité de la sonde d'intensité maximale et l'intensité maximale observée dans l'échantillon utilisé à la précédente étape. Cette différence sera ajoutée par la suite à l'intensité normalisée maximale de l'échantillon pour créer une nouvelle mesure normalisée maximale.

Dans le second cas, on utilise le même principe mais en utilisant les valeurs des intensités minimales.

Si après normalisation une intensité est négative ou nulle, alors lui est attribuée comme valeur celle du bruit de fond définie pour un individu donné de la façon suivante : Le bruit de fond correspond à la moyenne de la médiane des intensités des sondes de contrôles négatifs pour l'état méthylé et de la médiane des intensités des sondes de contrôles négatifs pour l'état non méthylé. Une fois les intensités des signaux méthylés et non méthylés normalisées par la méthode "SWAN", les β sont calculés selon la méthode habituelle. La méthode "SWAN" permet de rendre la distribution des échantillons identique, mais la distribution des intensités des sondes Infinium I est encore très différente de la distribution des intensités des sondes Infinium II car le nombre de sondes Infinium II est supérieur au nombre de sondes Infinium I.

La figure 3.29 montre la fonction de densité des β obtenue après normalisation par la méthode "SWAN".

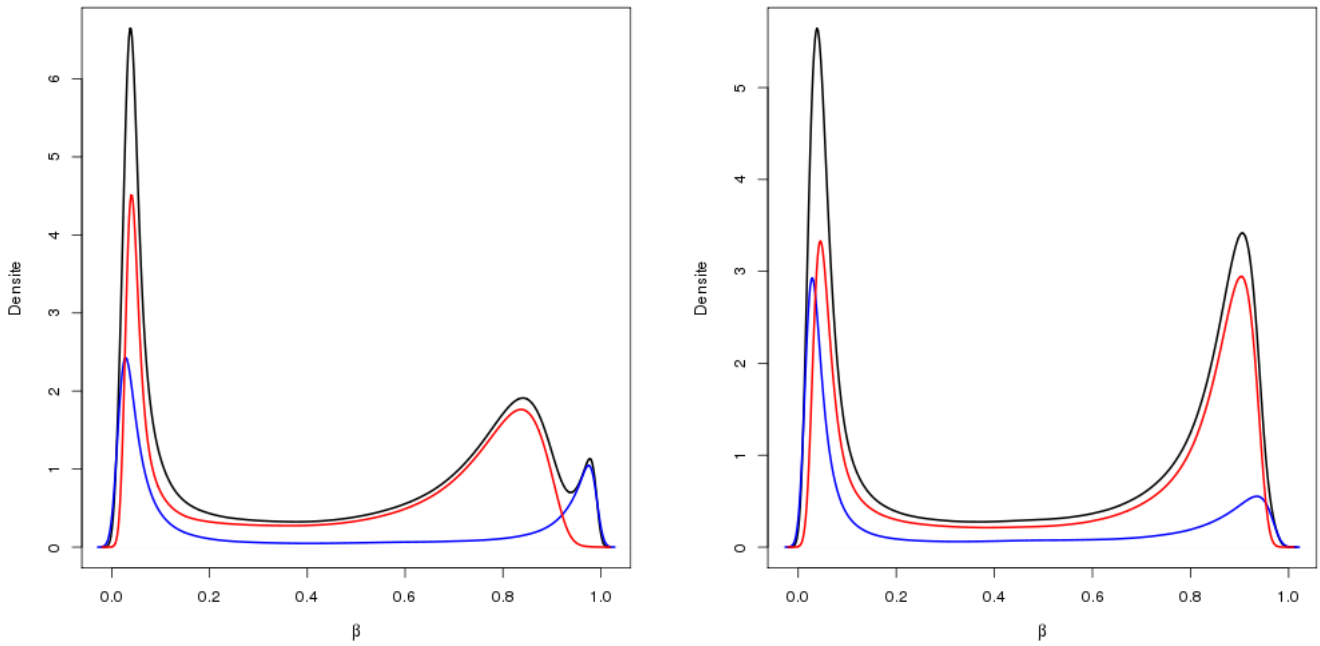

FIGURE 3.29 – Fonction de densité des β brutes (gauche) et normalisées par la méthode "SWAN" (droite). La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

On observe sur la figure 3.29 que les courbes bleues et rouges ont des pics qui ne sont plus séparés comme c'était le cas sur les données brutes. La méthode "SWAN" corrige donc bien le biais du type de sonde.

Méthode "BMIQ"

La méthode proposée par TESCHENDORFF, MARABITA *et al.* 2013 appelée "Beta MIxture Quantile dilation" est une méthode de normalisation basée sur un modèle de mélanges (*"mixture model"* en anglais). Un état de méthylation (non méthylé, hémiméthylé ou méthylé) est attribué à chaque sonde par le critère de probabilité maximale.

On crée un modèle de mélange composés de trois distributions bêta pour les trois états (non méthylé "U", hémiméthylé "H" et méthylé "M") par type de sonde. Une distribution bêta est définie par deux paramètres de forme notés habituellement α et β, pour plus de lisibilité ils seront respectivement renommés dans ce travail a et b. Le modèle de mélange a pour paramètres pour les trois distributions bêta des sondes Infinium I :

$$\{(a^I_U, b^I_U), (a^I_H, b^I_H), (a^I_M, b^I_M)\}$$

Et pour les trois distributions bêta des sondes Infinium II :

$$\{(a_U^{II}, b_U^{II}), (a_H^{II}, b_H^{II}), (a_M^{II}, b_M^{II})\}$$

Pour les sondes de type Infinium II non méthylées (U), on transforme leurs probabilités d'appartenance à l'état non méthylé en quantile en utilisant l'inverse de la distribution bêta cumulative avec pour paramètres bêta (a_U^I, b_U^I) estimés à partir de la distribution des sondes de type Infinium I non méthylées. Les valeurs normalisées des sondes de type Infinium II non méthylées sont notées η_U^{II}. Pour les sondes de type Infinium II méthylées (M), on transforme leurs probabilités d'appartenance à l'état méthylé en quantile en utilisant l'inverse de la distribution bêta cumulative avec pour paramètres bêta (a_M^I, b_M^I) estimés à partir de la distribution des sondes de type Infinium I méthylée. Les valeurs normalisées des sondes de type Infinium II méthylée sont notées η_M^{II}. Pour les sondes de type Infinium II hémiméthylées (H), on réalise une transformation de dilatation (de l'échelle) pour adapter les données dans l'intervalle avec les extrémités définit par $max\{\eta_U^{II}\}$ et $min\{\eta_M^{II}\}$.

Voici en détail comment procéder à la normalisation :

Pour la première étape qui consiste à attribuer les sondes à un des groupes (U, H ou M), on modélise chaque β ainsi :

$$p(\beta^t) = \pi_U^t B(\beta|a_U^t, b_U^t) + \pi_H^t B(\beta|a_H^t, b_H^t) + \pi_M^t B(\beta|a_M^t, b_M^t)$$

où B est une densité de probabilité bêta et t désigne le type de la sonde Infinium (I ou II). Les paramètres (π, a, b) sont déduits en utilisant l'algorithme espérance-maximisation (*EM* pour *Expectation Maximisation* en anglais) décrit par Y. Ji *et al.* 2005. Les paramètres estimés seront notés ainsi (π_s^t, a_s^t, b_s^t) où t désigne le type de la sonde et s un des trois états (U, H ou M). Les moyennes résultantes des distributions bêta estimées sont notées : m_s^t et calculées ainsi :

$$m_s^t = \frac{a_s^t}{a_s^t + b_s^t}$$

On définit U_{II}, H_{II} et M_{II} comme étant l'ensemble des sondes pour les états méthylés, hémiméthylés et non méthylés d'après le critère de probabilité maximale. On définit respectivement U_{II}^L et U_{II}^R comme l'ensemble des sondes U_{II} avec des valeurs β inférieures et supérieures à m_U^{II}. Idem pour les sondes méthylées avec M_{II}^L et M_{II}^R comme l'ensemble des sondes M_{II} avec des valeurs β inférieures

et supérieures à m_M^{II}. La séparation au niveau de la moyenne est nécessaire car les probabilités d'appartenance à un état de méthylation estimées à partir de l'algorithme EM sont bilatérales.

Ensuite pour la deuxième étape qui consiste à normaliser les sondes Infinium II non méthylées, on estime pour les sondes U_{II}^L la probabilité d'appartenir à l'état non méthylé :

$$p = P(U|\beta_{U_{II}^L}) = F(\beta_{U_{II}^L}|a_U^{II}, b_U^{II})$$

où F est la fonction de distribution bêta cumulative. Ensuite il faut transformer ces probabilités en quantiles (ce qui correspond aux valeurs β) mais en utilisant les paramètres des types I c'est-à-dire :

$$q = F^{-1}(p|a_U^I, b_U^I)$$

Les valeurs obtenues correspondent aux valeurs β normalisées ($\eta_{U_{II}^L} = q$). Une transformation identique est réalisée pour les sondes U_{II}^R mais en prenant $1 - F$ à la place de F.

On réalise la même procédure avec les sondes M_{II}^L et M_{II}^R pour normaliser les sondes Infinium II méthylées.

Et enfin pour normaliser les sondes Infinium II hémiméthylées, nous pouvons utiliser une approche empirique car leur distribution est prise en sandwich entre les distributions des sondes non méthylées et méthylées. L'approche empirique permet de contourner le problème de la mauvaise description des sondes hémiméthylées par une distribution bêta. Il faut d'abord déterminer le minima ($minH = min\beta_H^{II}$) et le maxima ($maxH = max\beta_H^{II}$) des sondes hémiméthylées de type 2, puis on défini $\Delta_H^{(\beta)} = maxH - minH$. On détermine ensuite, le minima des sondes méthylées de type 2 ($minM = min\beta_M^{II}$) et le maxima des sondes non méthylées de type 2 ($maxU = max\beta_U^{II}$). Ces extrema représentent des valeurs solides, car ils ne représentent pas des extrema de la borne $[0; 1]$. En effet, les valeurs $maxU$, $minH$, $maxH$ et $minM$ ne sont pas proches de 0 ou 1.

Il faut ensuite déterminer les distances : $\Delta_{UH} = minH - maxU$ et $\Delta_{HM} = -maxH + minM$. On calcule ensuite les valeurs normalisées minimales ($nminH = max\{\eta_U^{II}\} - \Delta_{UH}$) et maximales ($nmaxH = min\{\eta_M^{II}\} - \Delta_{HM}$) pour les sondes hémiméthylées. Les valeurs β normalisées pour les sondes hémiméthylées sont obtenues par une transformation conforme (décalage et dilatation), c'est-à-dire :

$$\eta_H^{II} = nminH + d_f(\beta_H^{II} - minH)$$

où $d_f = \Delta_H^{(\eta)} / \Delta_H^{(\beta)}$ est le facteur de dilatation.

Il est important de noter que la transformation conforme implique un changement d'échelle non uniforme des valeurs β des sondes hémiméthylées car elle dépend de la valeur β de la sonde, ceci est nécessaire pour éviter l'apparition de "trou" dans la distribution normalisée.

La figure 3.30 montre la fonction de densité des β obtenue après normalisation par la méthode "BMIQ".

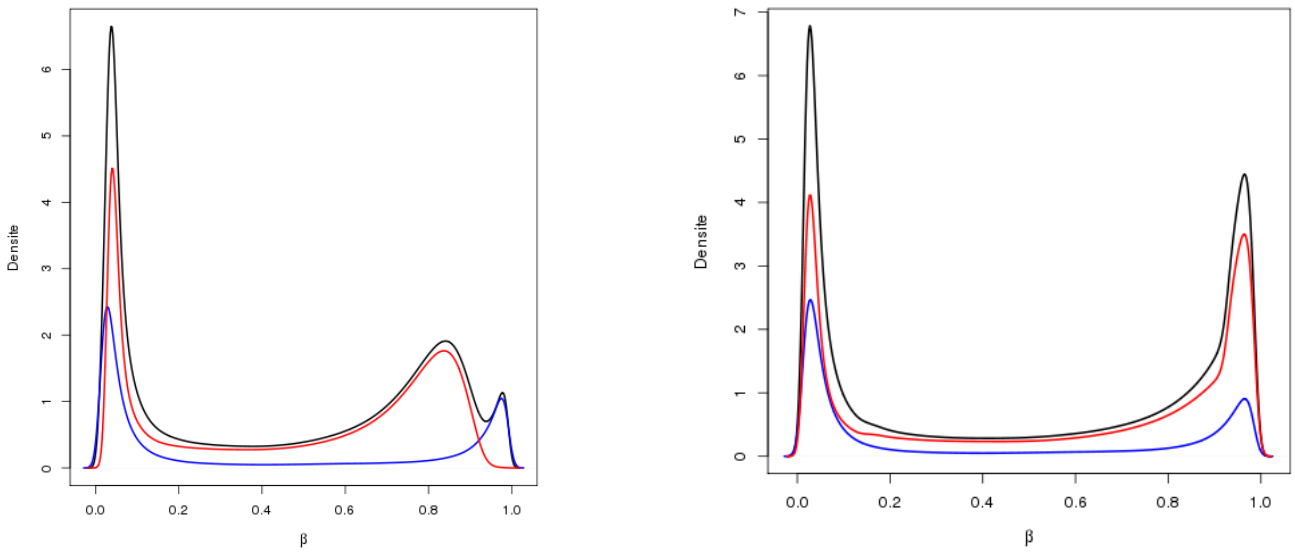

FIGURE 3.30 – Fonction de densité des β brutes (gauche) et normalisées par la méthode "BMIQ" (droite). La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

On observe sur la figure 3.30 que les courbes bleues et rouges ont des pics qui ne sont plus séparés comme c'était le cas sur les données brutes. La méthode "BMIQ" corrige donc bien le biais du type de sonde.

8.3 Correction du biais du fluorochrome

Méthode d'Illumina

Cette méthode de normalisation utilise les 186 sondes de contrôles de normalisation. Ces sondes sont conçues pour cibler la même région dans des gènes de ménage et ne comprennent pas de sites CpG. Les gènes de ménage sont des gènes qui sont exprimés à des taux constants (sans mécanisme de régulation) dans toutes les cellules d'un organisme car le produit de leur expression est indispensable au métabolisme de base des cellules. Les intensités obtenues par ces sondes doivent donc être équiva-

lentes quel que soit l'échantillon utilisé. Il s'agit de contrôle réalisé par échantillon. La performance des contrôles est surveillée dans le rouge pour les nucléotides A et T et dans le vert pour les nucléotides C et G. On calcule par individu la valeur moyenne des intensités des contrôles de normalisation séparément dans le vert (noté : $Moyenne_{vert}$) et dans le rouge (noté : $Moyenne_{rouge}$). On définit une valeur indice par individu qui correspond à la moyenne des deux valeurs moyennes précédemment calculées ($Moyenne_{vert}$ et $Moyenne_{rouge}$). On sélectionne une valeur indice d'un échantillon comme valeur de référence pour normaliser les autres échantillons. La méthode de sélection de la valeur de référence parmi les valeurs indices qu'Illumina utilise dans son logiciel GenomeStudio étant inconnue, les packages Bioconductor "methylumi" et "minfi" laissent le choix à l'utilisateur de sélectionner l'échantillon servant de référence. Toutefois, ils proposent différents choix par défaut : le package "minfi" sélectionne comme référence le premier échantillon du jeu de données alors que le package "methylumi" sélectionne comme référence l'échantillon dont la valeur $\frac{Moyenne_{rouge}}{Moyenne_{vert}} - 1$ est la plus proche de 0. On divise ensuite toutes les moyennes rouges ($Moyenne_{rouge}$) et les moyennes vertes ($Moyenne_{vert}$) par la valeur de référence, on obtient des valeurs factrices rouge et verte. En multipliant respectivement les valeurs factrices verte et rouge d'un échantillon aux intensités vertes et rouges des sondes de mesures de ce même échantillon, nous obtenons les valeurs normalisées des intensités des sondes de mesures pour cet échantillon.

La figure 3.31 représente les 588 fonctions de densités des logarithmes binaires des intensités Infinium II pour les états méthylés (verts) et non méthylés (rouges). Nous pouvons voir sur la figure de gauche (avant normalisation) que les fonctions de densités vertes et rouges ne sont pas superposables alors qu'elles le sont après normalisation (à droite).

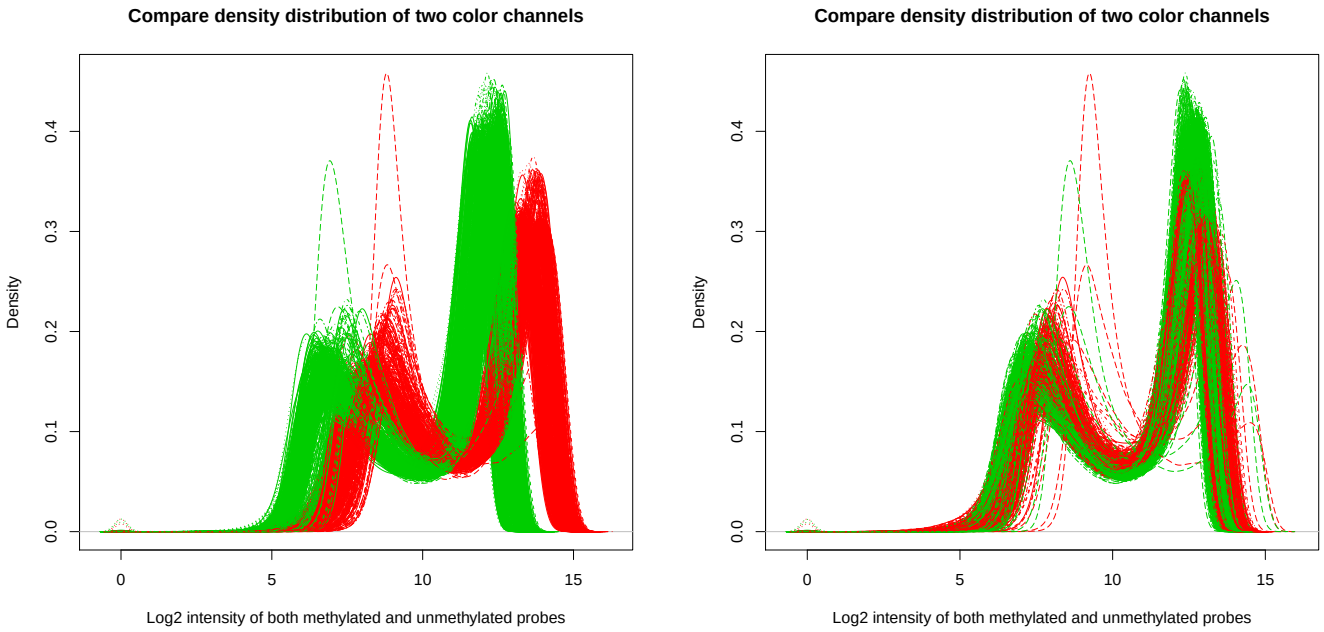

FIGURE 3.31 – Fonction de densité des 588 logarithmes binaires des intensités Infinium II pour les états méthylés (verts) et non méthylés (rouges). Avant correction à gauche et après correction à droite.

On voit sur la figure 3.32 qu'après normalisation du biais du fluorochrome (à droite), le biais lié à l'utilisation de deux types de sondes est atténué sans toutefois être totalement corrigé.

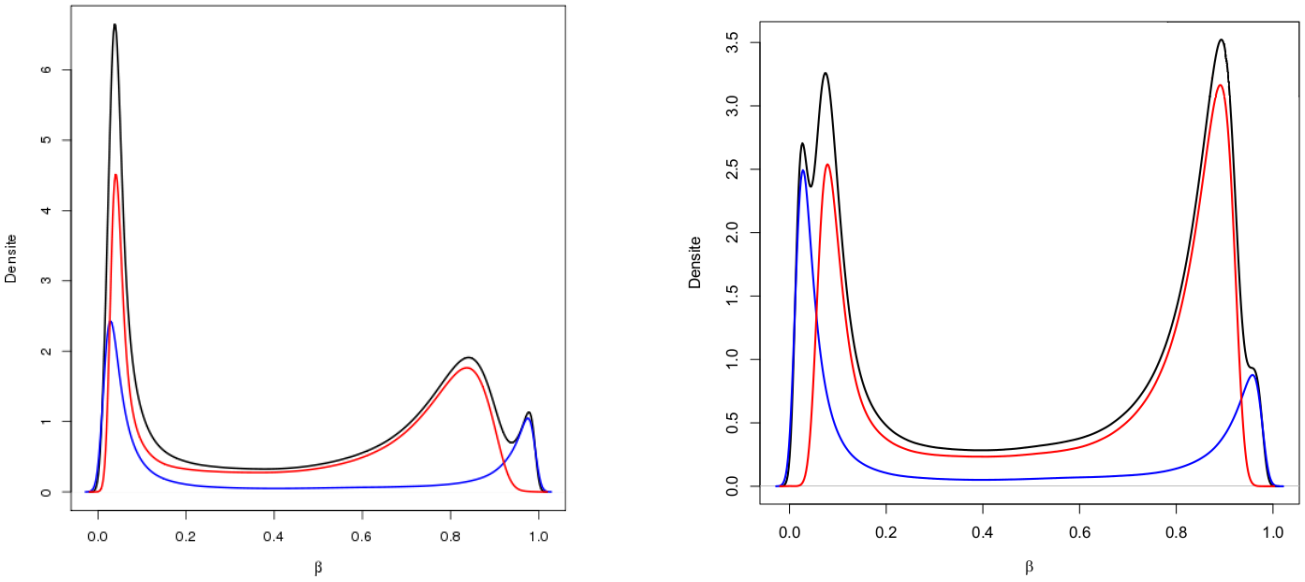

FIGURE 3.32 – Fonction de densités des β non normalisés (à gauche) et normalisés avec la méthode d'Illumina (à droite). La courbe bleue représente les valeurs obtenues par Infinium I, la courbe rouge représente les valeurs obtenues par Infinium II et la courbe noire représente les valeurs obtenues par Infinium I & II.

8.4 Correction de l'effet de lot

Méthode "ABnorm"

La normalisation par quantile rend les distributions des intensités identiques entre les 588 échantillons. Cela a pour conséquence de supprimer la différence de variabilité entre les différents lots. En effet, si tous les échantillons ont la même distribution d'intensités, alors les lots ont également la même distribution d'intensités. L'objet de la normalisation par quantile est de minorer les variations artefactuelles (c'est-à-dire les variations liées à l'effet de lot) et de ne faire ressortir que les variations biologiques.

La correction "ABnorm" se différencie de la correction "SWAN" car "ABnorm" est une normalisation par quantile des intensités entre les échantillons alors que "SWAN" est une normalisation par quantile entre les intensités Infinium I et Infinium II au sein d'un même échantillon. La méthode "ABnorm" testée ici est basée sur une méthode proposée à l'origine pour la biopuce Illumina HM27k (Z. Sun, Chai *et al.* 2011). Je l'ai adaptée ici pour la puce HM450k en normalisant par quantile les intensités des signaux méthylés (signal B) et non méthylés (signal A), séparément pour les sondes de type Infinium I et II. Le principe de la normalisation par quantile est de trier par ordre croissant les intensités, puis d'attribuer pour chaque rang la valeur correspondante à la moyenne des valeurs du rang. Une fois les nouvelles valeurs obtenues, il faut réordonner les valeurs dans leur ordre initial. Une normalisation par quantile est réalisée pour les intensités B de type Infinium I, B de type Infinium II, A de type Infinium I et A de type Infinium II séparément. Les intensités des 588 échantillons ont une distribution identique au sein de chacun des 4 groupes. Pour illustrer cette méthode, prenons les intensités des signaux méthylés. Pour chacun individu, les intensités sont triées par ordre croissant. L'intensité d'un rang i pour un individu ne correspondra pas au même site CpG que l'intensité du même rang pour un autre individu. La moyenne de toutes les valeurs du rang i remplacera les valeurs de ce rang. Les individus auront donc la même valeur pour le rang i mais cela ne correspondra pas au même site CpG. Lorsque les normalisations par quantile ont été réalisées, le β est ensuite calculé comme précédemment.

La normalisation par quantile des intensités va rendre les distributions des intensités au sein des individus identiques ce que confirme la figure 3.33. Cette figure représente des diagrammes en boites des intensités des signaux méthylés non normalisés (gauche) et normalisés par la méthode "ABnorm" (droite) par puces (ou lots) pour les 588 échantillons. On voit sur la figure de droite que les différents

lots ont la même distribution d'intensités.

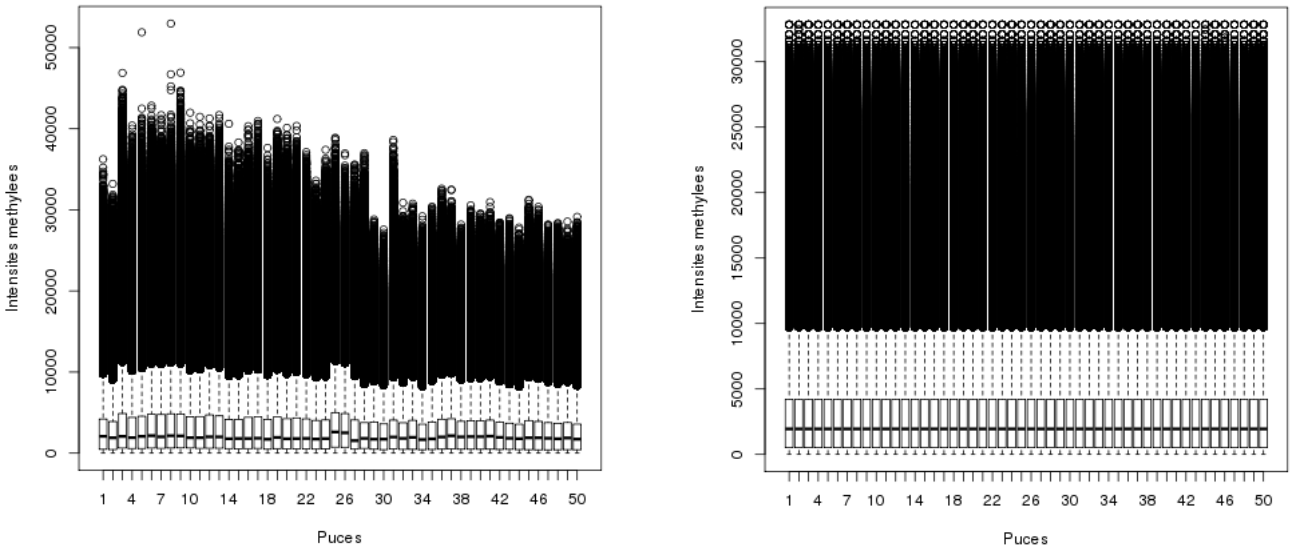

FIGURE 3.33 – Diagrammes en boites des intensités des signaux méthylés bruts (gauche) et normalisés par la méthode "ABnorm" (droite) par puces (ou lots) pour les 588 échantillons.

Le fait de rendre les distributions des intensités identiques va se répercuter sur les distributions des valeurs des β, comme on le voit sur les diagrammes en boites de la figure 3.34. La figure 3.34 représente des diagrammes en boites des β calculées à partir des valeurs des intensités non normalisées (gauche) et à partir des valeurs des intensités par la méthode "ABnorm" (droite) par lots (ou puces) pour les 588 échantillons. La distribution des β n'est pas identique entre les lots après normalisation mais on remarque tout de même une uniformisation et donc une diminution de l'effet de lot.

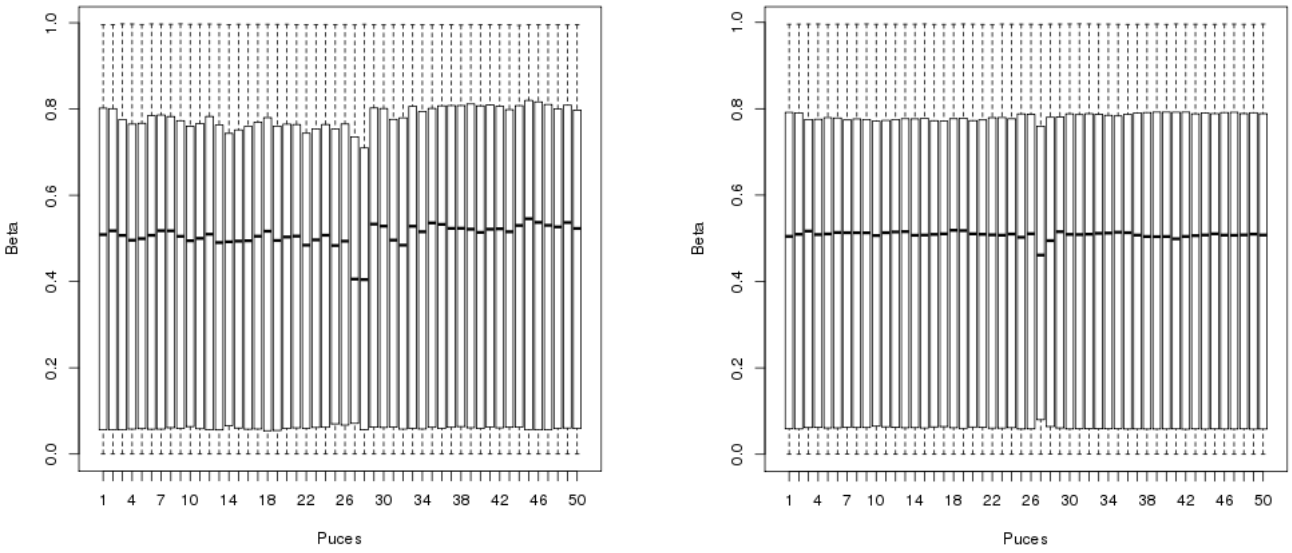

FIGURE 3.34 – Diagrammes en boites des β calculées à partir des valeurs des intensités non normalisées (gauche) et à partir des valeurs des intensités par la méthode "ABnorm" (droite) par puces (ou lots) pour les 588 échantillons.

Dans l'optique d'améliorer la méthode de normalisation "ABnorm", j'ai proposé une variante de la méthode "ABnormNeg" qui inclut les sondes de contrôles négatifs lors de la normalisation par quantile des intensités. Ces intensités sont traitées comme s'il s'agissait d'intensités de sondes mesurant les niveaux de méthylation. Les graphiques représentant les fonctions des densités des intensités, les diagrammes en boites des intensités du signal méthylé et des β étant identiques à ceux obtenus avec la méthode "ABnorm", ne sont pas présentés dans ce rapport.

Autres méthodes de normalisation par quantile

La méthode "Lumi" consiste à normaliser par quantile non pas les intensités méthylées et non méthylées comme dans la méthode "ABnorm", mais les intensités d'une part obtenues via le fluorochrome "Cy3" et d'autre part les intensités obtenues via le fluorochrome "Cy5" (Z. Sun, Chai *et al.* 2011). Cette normalisation est identique à "ABnorm" pour les sondes Infinium II où chaque état est mesuré par un fluorochrome spécifique mais diffère pour les sondes Infinium I où chaque état peut être mesuré soit par le fluorochrome Cy3 ou Cy5.

La méthode "QNBeta" consiste à normaliser par quantile les valeurs β (Pidsley *et al.* 2013 ; Z. Sun, Chai *et al.* 2011). Cette méthode a l'inconvénient d'être trop drastique et d'effacer certaines différences biologiques pertinentes.

Méthode "ComBat"

La méthode largement utilisée "ComBat" est proposée par Johnson *et al.* 2007 et est disponible dans le package Bioconductor "sva". Cette méthode utilise une approche empirique bayésienne paramétrique ou non paramétrique pour corriger l'effet de lot en faisant l'hypothèse que l'effet de lot est composé d'un effet additif suivant une loi normale et d'un effet multiplicatif suivant une loi gamma inverse. Pour plus de détails sur la méthode statistique de "ComBat", j'invite le lecteur à se référer à l'article la décrivant en détail (Johnson *et al.* 2007).

Les diagrammes en boites de la figure 3.35 représentent les distributions des valeurs β normalisées (droite) et non normalisées (gauche) par la méthode "ComBat" et par puces.

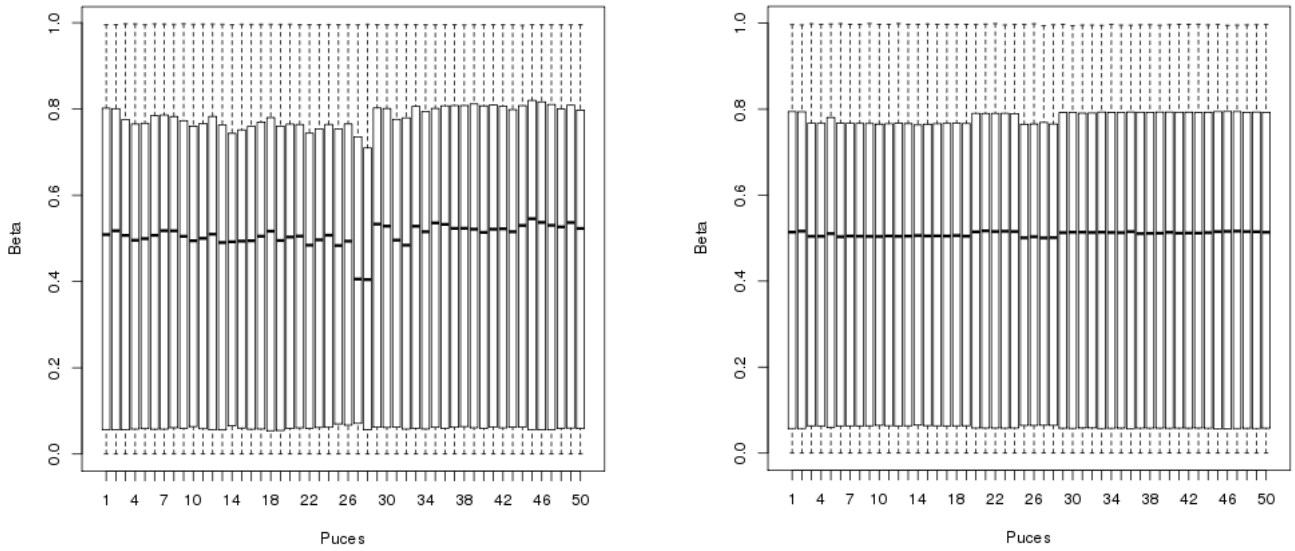

FIGURE 3.35 – Diagrammes en boites des β non normalisées (à gauche) et normalisées (à droite) par la méthode ComBat pour les 588 échantillons.

On peut voir qu'après la normalisation les distributions des valeurs β sont plus semblables qu'avant la normalisation, la normalisation par la méthode "ComBat" a diminué la variabilité entre les puces et donc l'effet de lot.

Autres méthodes

Il existe diverses autres méthodes pour corriger l'effet de lot comme les méthodes "LOESS", "LOWESS" et "SPLINES" qui n'ont pas été traitées dans ce travail de thèse mais qui mériteraient d'être étudiées dans le cas des données de méthylation.

8.5 Combinaison de méthodes

Les méthodes présentées ci-dessus ne corrigent qu'un type de biais, or les données générées par la puce HM450k sont soumises à plusieurs biais. Il est donc nécessaire de combiner, c'est-à-dire d'appliquer successivement, plusieurs méthodes corrigeant différents biais présents. Par exemple, l'une des combinaisons possibles est d'appliquer la méthode "Noob" pour corriger le bruit de fond, puis d'appliquer la méthode d'Illumina pour corriger le biais du fluorochrome et enfin d'appliquer la méthode "SWAN" pour corriger le biais lié à l'utilisation de deux types de sondes.

Les différentes méthodes corrigent les biais présents plus ou moins bien. Il est donc nécessaire de les comparer pour évaluer lesquelles sont les plus pertinentes pour nos données. La partie suivante traite de la façon dont ont été comparées les méthodes ainsi que les combinaisons de méthodes.

9 Comparaison des méthodes

Pour déterminer l'efficacité des différentes méthodes et pour pouvoir les comparer, j'ai utilisé les 11 réplicats techniques. Il s'agit de 11 individus qui ont été mesurés deux fois dans des lots (ou puces) différents, ce qui permet de juger de la qualité et de la reproductibilité des mesures et servir ainsi de référence pour estimer les variations dues à des conditions expérimentales différentes. L'hypothèse sous-jacente est qu'après normalisation, les deux mesures du même individu doivent avoir une meilleure concordance.

Pour mesurer la concordance entre les deux mesures, j'ai utilisé le coefficient de concordance de Lin (Lin 1989), le coefficient de détermination R^2 (il s'agit du coefficient de corrélation de Pearson mis au carré, bien que moins adapté que le coefficient de concordance car plus spécifique pour mesurer la corrélation, il permet tout de même de donner une idée de reproductibilité des résultats). Pour chaque réplicat indépendamment, j'ai calculé le coefficient de concordance de Lin et le coefficient de détermination R^2 entre les valeurs β de la première et de la seconde mesure. Plus les valeurs du coefficient de concordance et de détermination sont importantes, idéalement égales à 1, plus la reproductibilité entre les deux mesures est bonne et donc plus efficace est la normalisation.

J'ai également réalisé un test de Kolmogorov-Smirnov pour comparer les fonctions de distribution des valeurs β entre les deux mesures après normalisation indépendamment pour les 11 réplicats. On considère ici que plus la statistique de test est faible plus les deux distributions sont semblables et donc plus efficace est la normalisation.

J'ai ensuite réalisé un graphique de type Bland-Altman (Altman & Bland 1983) permettant de représenter les différences entre deux mesures (voir figure 3.36). La moyenne entre deux mesures est représentée sur l'axe des abscisses et la différence entre deux mesures est sur l'axe des ordonnées. Dans le cas idéal où les deux mesures seraient totalement concordantes, on aurait tous les points alignés sur la droite d'ordonnée 0.

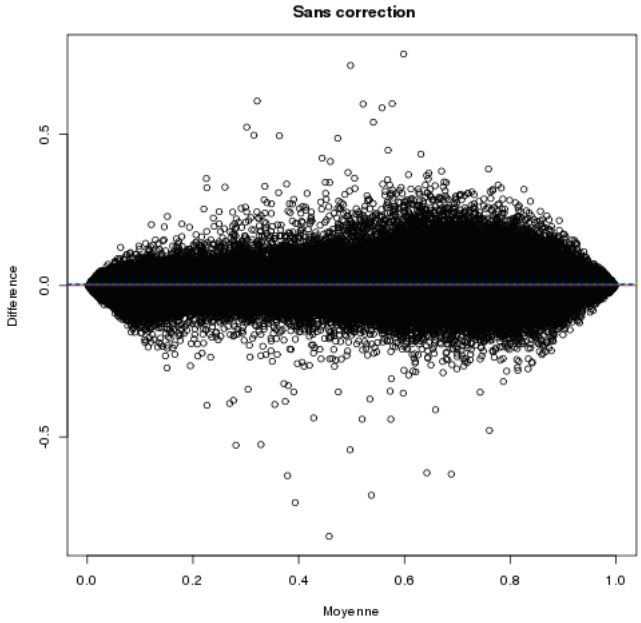

FIGURE 3.36 – Exemple de graphique de type Bland-Altman représentent les valeurs β non corrigées des deux mesures d'un même individu.

La moyenne des différences est également calculée, ce qui permet d'avoir une valeur pouvant définir la variabilité entre deux mesures d'un même individu, idéalement cette valeur serait égale à 0. J'ai également réalisé un graphique avec pour axe chacune des deux mesures (voir figure 3.37). Ici dans le cas idéal, les points seraient alignés sur la droite diagonale partant de l'origine ($x = 0, y = 0$) et allant au point ($x = 1, y = 1$).

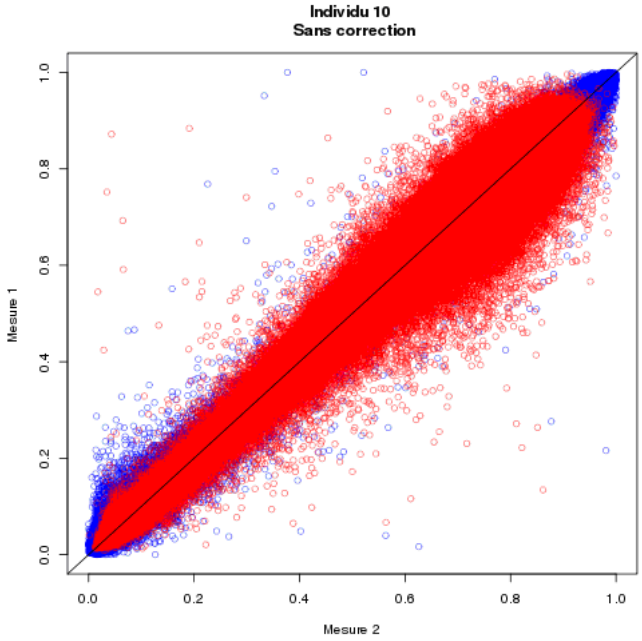

FIGURE 3.37 – Exemple de graphique représentent les valeurs β non corrigées des deux mesures d'un même individu. Les points bleus représentent les β obtenus avec les sondes Infinium I et les points rouges ceux obtenus avec les sondes Infinium II.

Comme il s'avère que les sondes Infinium II sont moins précises et reproductibles que les sondes Infinium I, j'ai représenté l'écart type entre deux mesures par des diagrammes en boites en fonction du type de sonde utilisée (voir figure 3.38). Ces graphiques permettent d'avoir une représentation visuelle de la reproductibilité des deux types de sondes. La moyenne des écarts types est également calculée en fonction du type de sonde, ce qui permet d'avoir une valeur définissant la reproductibilité pour les deux types de sondes (idéalement à 0).

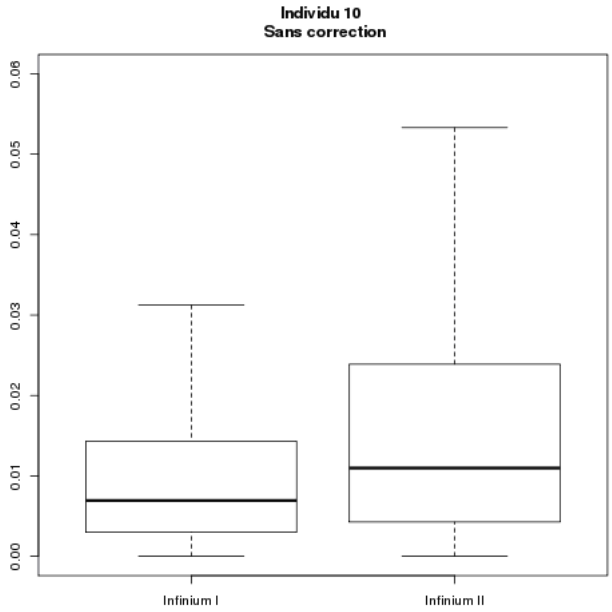

FIGURE 3.38 – Exemple de diagrammes en boites des écarts types entre les deux mesures des valeurs β non corrigées en fonction du type de sonde Infinium utilisé.

Tous ces critères permettent de mesurer la diminution des variations techniques dûe à la normalisation.

J'ai donc comparé les différentes méthodes de correction du bruit de fond, qu'elles soient basées sur la soustraction du bruit de fond (dont la méthode proposée par Illumina) ou sur la déconvolution (comme la méthode "Noob"). Via les différents critères utilisés, j'ai pu identifier la méthode "Noob" comme la plus efficace et donnant des résultats les plus concordants entre les deux mesures des réplicats. Pour le biais du fluorochrome, j'ai sélectionné la seule méthode disponible et proposée par défaut par Illumina car elle fonctionne correctement. En ce qui concerne le biais lié à l'utilisation de deux types de sondes, j'ai exclu la méthode "Peak-based Correction" car bien qu'efficace pour corriger le biais concerné, elle introduisait un nouveau biais pour les valeurs β proches de 0,5. J'ai comparé les diverses autres méthodes de correction et sélectionné la méthode "SWAN" selon les critères définis ci-dessus. En ce qui concerne l'effet de lot, j'ai choisi de ne pas le corriger par des méthodes standard, c'est-à-dire par des transformations de données, mais de l'intégrer dans les analyses statistiques en tant que variable d'ajustement dans les modèles de régression (voir le chapitre sur les analyses statistiques des associations méthylation-phénotype, p. 92). Cela a pour effet qu'il n'introduit plus de biais dans les résultats et d'éviter une nouvelle transformation des données tout en permettant d'estimer l'effet de chaque lot sur les varibales d'intérêts. Le biais lié à la contamination a été traité selon le même

procédé, c'est-à-dire que les différentes proportions des leucocytes ont été ajoutées comme variables d'ajustements. Le biais dû au taux de GC dans la sonde est un biais actuellement très peu étudié dans le cadre des données de la puce HM450k et n'est pas pris en compte dans ce travail de thèse.

J'ai ensuite comparé différentes combinaisons des méthodes sélectionnées ci-dessus. En examinant les différents critères ci-dessus, j'ai sélectionné la combinaison des méthodes "Noob", de la méthode d'Illumina pour corriger le biais du fluorochrome et de la méthode "SWAN".

10 Conclusions Correction & Normalisation

La première étape de mon travail de thèse était d'effectuer le contrôle qualité des données générées par la puce HM450k. J'ai dû exclure les échantillons ayant des valeurs aberrantes, que ce soit les individus ou les sondes de la puce. Ensuite, il a fallu caractériser les biais générés par l'utilisation de la puce puis trouver des méthodes pour les corriger. Pour cette dernière étape, j'ai combiné plusieurs méthodes pour corriger les différents biais (figure 3.39). La méthode Noob pour corriger le bruit de fond, suivie de la méthode proposée par Illumina pour corriger le biais du fluorochrome et enfin la méthode "SWAN" pour le biais lié aux deux types de sonde Infinium.

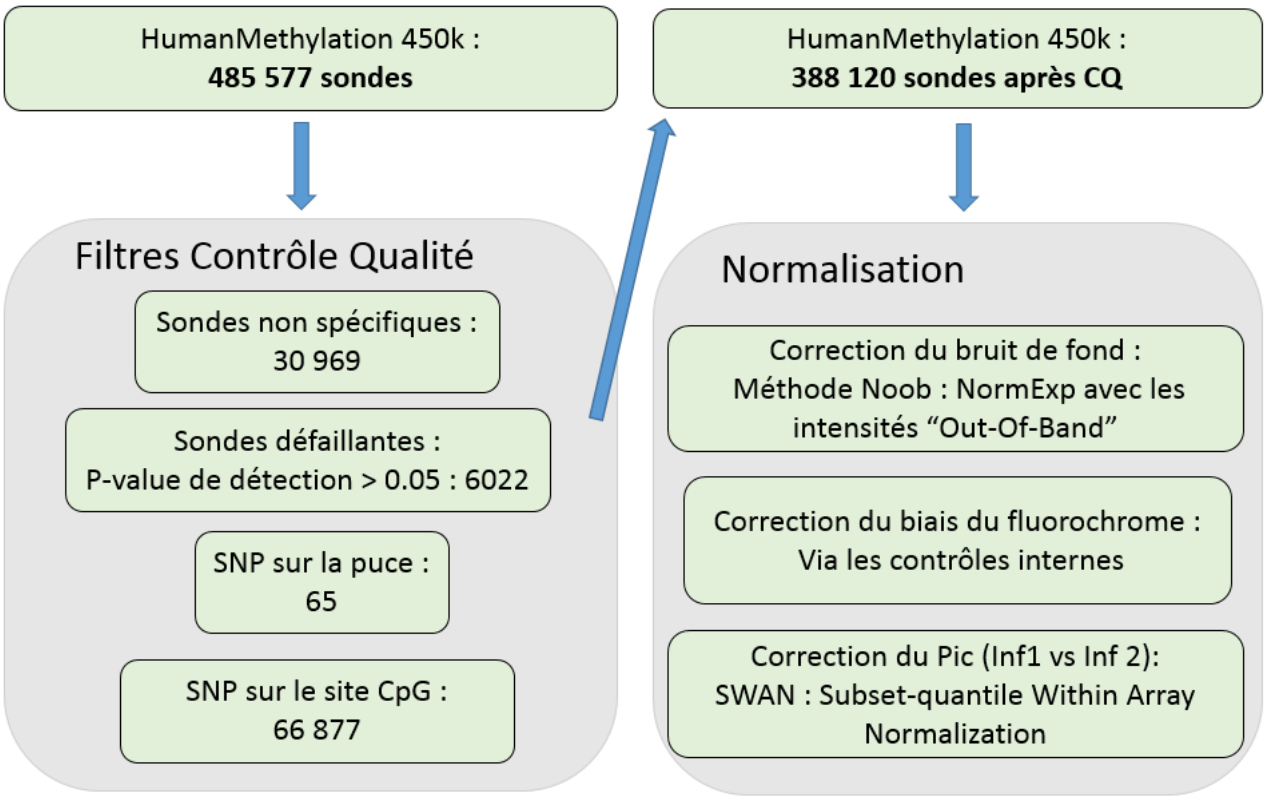

FIGURE 3.39 – Diagramme de flux du contrôle qualité et de la normalisation pour les études MARTHA et F5L-Pedigrees.

11 Validation des données

Pour valider la cohérence des données ainsi obtenues, j'ai étudié deux associations robustes déjà connues avec le niveau de méthylation. Tout d'abord, la méthylation du locus *F2RL3* qui est connue pour être plus faible chez les fumeurs que chez les non fumeurs (BREITLING *et al.* 2011 ; TSAPROUNI *et al.* 2014 ; ZEILINGER *et al.* 2013). Cette association est retrouvée dans l'étude MARTHA comme on peut le voir sur la figure 3.40 qui représente via des diagrammes en boîte les niveaux de méthylation du site CpG cg03636183 du gène *F2RL3* en fonction de la consommation de tabac. Sur cette figure l'on voit que les niveaux de méthylation sont plus faibles pour les fumeurs (la moyenne du β est de 0,65) que pour les non fumeurs (la moyenne du β est aux alentours de 0,8) avec un niveau intermédiaire pour les anciens fumeurs, ce qui permet de supposer que l'effet du tabac sur la méthylation du site CpG cg03636183 est temporaire.

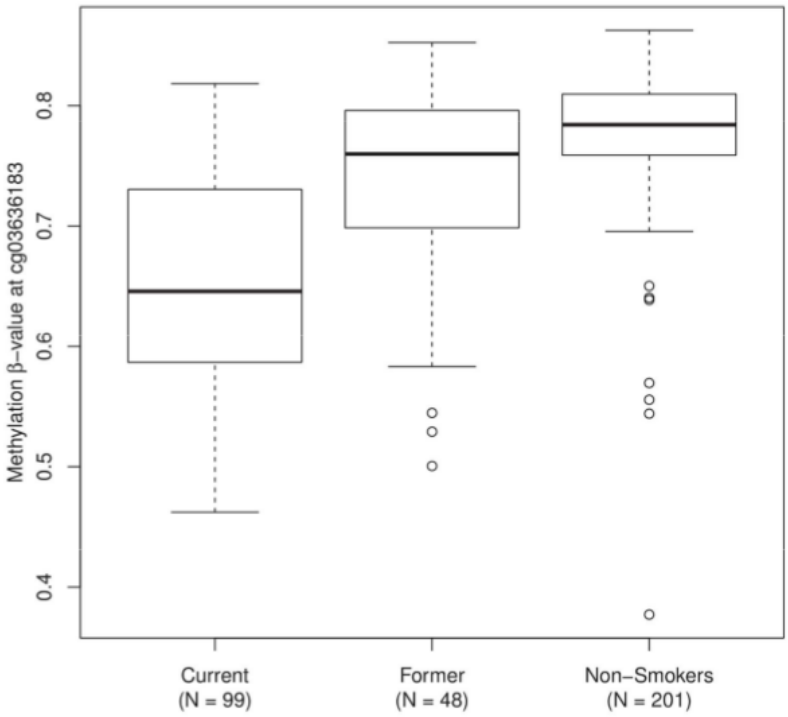

FIGURE 3.40 – Association entre le tabagisme et la méthylation du locus *F2RL3* (site CpG cg03636183) dans l'étude MARTHA.

On retrouve également dans l'étude MARTHA l'association entre le polymorphisme rs713586 et la méthylation du locus *DNAJC27/ADCY3* (GRUNDBERG *et al.* 2013). Comme nous pouvons le voir sur la figure 3.41, l'allèle T du polymorphisme rs713586 est associé à une plus faible méthylation du locus *DNAJC27/ADCY3*. Les individus homozygotes pour l'allèle T ont une méthylation moyenne d'environ 0,2, alors qu'elle est de 0,3 pour les hétérozygotes et d'environ 0,38 pour les homozygotes pour l'allèle C.

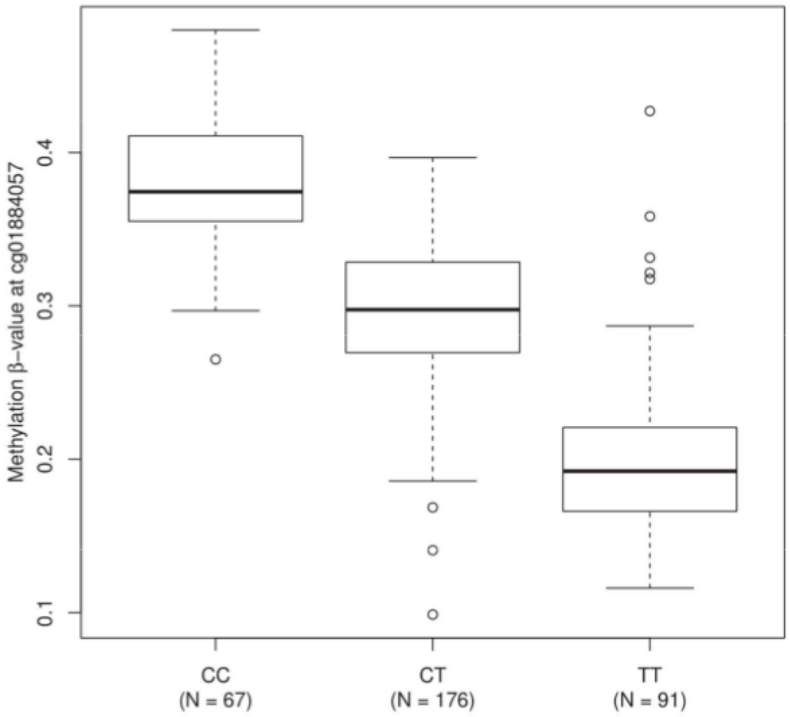

FIGURE 3.41 – Association entre le polymorphisme rs713586 et la méthylation du locus *DNAJC27/ADCY3* (site CpG cg01884057) dans l'étude MARTHA.

Ces résultats nous permettent de confirmer la validité de nos données ainsi que celle du contrôle qualité et de la normalisation.

Troisième partie

Étude à grande échelle de la méthylation de l'ADN

Chapitre 4

Étude d'association méthylome entier : pas à pas

La réalisation d'une étude d'association méthylome entier ("MWAS") se fait en plusieurs étapes quel que soit le type cellulaire ou le tissu analysé. La première étape d'une étude MWAS consiste à mesurer à proprement parler le méthylome. Dans ce travail de thèse, le méthylome est mesuré en utilisant la puce HM450k (voir la partie précédente décrivant la puce HumanMethylation450k, p. 24). Une fois le méthylome mesuré, la deuxième étape consiste à réaliser un contrôle qualité et à normaliser les données générées par le scanner (voir la partie précédente décrivant le contrôle qualité et normalisation, p. 31). L'étape suivante, sujet de cette partie, est l'analyse statistique des données qui peut se décomposer ainsi :

- les tests d'association avec le phénotype d'intérêt ;
- la correction pour la multiplicité des tests réalisés ;
- la réplication, le cas échéant, dans d'autres échantillons indépendants. Pour confirmer les résultats trouvés dans un échantillon donné, il est nécessaire d'effectuer une réplication dans une étude indépendante, ce qui permettra de confirmer le lien statistique entre les deux variables et ainsi éviter que le hasard n'intervienne dans nos résultats.
  - et la méta-analyse des résultats obtenus au sein de différents échantillons ;
- la dernière étape consiste en l'interprétation biologique des résultats obtenus ainsi qu'en leur discussion.

1 Analyse statistique des associations méthylation-phénotype

Une fois l'étape du contrôle qualité effectuée, vient ensuite l'analyse statistique des données. Il est possible d'utiliser différentes méthodes pour réaliser une analyse MWAS, dont le choix dépend entre autres du type de données utilisées (individus provenant de famille ou sujets indépendants), de l'information a priori (approche bayésienne ou fréquentiste) ou également du phénotype étudié (variable biologique quantitative ou binaire comme la présence ou non d'une pathologie). Les méthodes standards utilisées dans les analyses génomiques sont basées sur les modèles linéaires généralisés, ce qui permet d'ajuster sur différents facteurs de confusions pertinents comme l'âge, le sexe, le tabagisme, etc. On utilisera un modèle linéaire lorsque la variable étudiée sera quantitative (par exemple une variable biologique comme le taux de cholestérol), alors que l'on utilisera un modèle logistique lorsque le phénotype étudié sera binaire. Il est également fréquent d'utiliser des modèles mixtes lorsque l'on dispose de données provenant de famille afin de prendre en compte la non indépendance des données familiales.

Dans le contexte de la méthylation, la variable définissant le niveau de méthylation peut être la variable expliquée (celle que l'on cherche à modéliser) ou une des variables explicatives (celles qui permettent d'expliquer la variable d'intérêt). Le choix se fera en fonction de la question à laquelle on essaie de répondre. Dans le cas où l'on cherche à expliquer une maladie ou un trait phénotypique quantitatif par des niveaux de méthylation, la variable représentant le niveau de méthylation sera une variable explicative. Tandis que si l'on cherche à expliquer le niveau de méthylation d'un site CpG, alors la variable définissant le niveau de méthylation sera la variable expliquée.

1.1 Modèle linéaire

Dans un modèle linéaire, la variable expliquée Y est influencée de manière linéaire par les variables explicatives X. Voici un exemple où nous expliquons la variable biologique quantitative d'intérêt Y par un modèle linéaire en fonction de l'âge, du sexe et de la méthylation d'un site CpG :

$$E(Y) = \alpha + \beta_1 * X_{age} + \beta_2 * X_{sexe} + ... + \beta_z * X_{CpG_n} + \epsilon$$

Avec :

— α : l'intercepte ou constante de régression ;

— β_1 : l'effet de l'âge sur la variable Y ;

— β_2 : l'effet du sexe sur la variable Y ;

— β_z : l'effet de la méthylation du site CpG n sur la variable Y ;

— ϵ : une erreur résiduelle traduisant la part de Y non expliquée par les variables X du modèle, en d'autres termes correspond à la variabilité biologique entre les individus. Un tel modèle linéaire repose sur l'hypothèse que cette erreur résiduelle suit une loi normale de moyenne 0 et de variance σ^2 estimée par la régression.

La régression va chercher à expliquer la variable Y par les différentes variables explicatives X en calculant leur effet β sur celle-ci. La méthode mathématique permettant de calculer les coefficients β ainsi que l'écart type associé est généralement l'estimation par le maximum de vraisemblance. C'est ce type de modèle linéaire que j'ai principalement utilisé dans mon travail de thèse lorsque je me suis intéressé à l'étude d'un trait quantitatif. Par exemple, lorsque j'ai étudié si des niveaux de méthylation étaient associés à l'indice de masse corporelle (IMC), j'ai utilisé le modèle suivant :

$$E(IMC) = \alpha + \beta_1 * X_{age} + \beta_2 * X_{sexe} + \beta_3 * X_{lymphocyte} + \beta_4 * X_{monocyte} + \\ + \beta_5 * X_{basophile} + \beta_6 * X_{eosinophile} + \beta_7 * X_{neutrophile} + \beta_8 * X_{CpG_n} + \epsilon$$

Lorsque le phénotype d'intérêt est binaire, le modèle logistique est généralement utilisé.

1.2 Modèle logistique

Dans le cas du modèle logistique, on modélise la probabilité qu'une variable binaire Y prenne la valeur 1 à partir de la formulation ci-après :

$$P(Y = 1) = \frac{e^{\alpha+\beta_1*X_{age}+\beta_2*X_{sexe}+...+\beta_z*X_{CpG_n}}}{1 + e^{\alpha+\beta_1*X_{age}+\beta_2*X_{sexe}+...+\beta_z*X_{CpG_n}}}$$

Cette formule permet de s'assurer que la valeur de la probabilité $P(Y = 1)$ soit bien comprise entre 0 et 1. L'utilisation de l'exponentielle dans le numérateur impose une limite inférieure de 0, tandis que le dénominateur impose une limite supérieure de 1.

Comme indiqué précédemment, ce type de modèle est couramment utilisé lorsque l'on souhaite modéliser le risque de survenue d'une maladie. Au cours de mon travail de thèse, j'ai également utilisé ce modèle pour déterminer si des profils de méthylation différaient entre les individus porteurs et non porteurs d'une mutation (voir la partie sur la recherche de facteurs épigénétiques pouvant expliquer la pénétrance incomplète du Facteur V de Leiden, p. 142). Dans ce cas, la variable Y ne traduit pas

la présence/absence d'une maladie mais la présence/absence d'une caractéristique génétique.

1.3 Modèle linéaire mixte

Le modèle linéaire simple présenté ci-dessous est un modèle à effets fixes :

$$E(Y) = \alpha + \beta_1 * X_{age} + \beta_2 * X_{sexe} + \beta_3 * X_{famille} + ... + \beta_z * X_{CpG_n} + \epsilon$$

Dans un modèle linéaire à effets mixtes, certains facteurs sont à effets fixes (ils interviennent au niveau de la moyenne du modèle) et d'autres sont à effets aléatoires (ils interviennent au niveau de la variance du modèle). Le fait de rajouter une variable aléatoire permet de prendre en compte la non indépendance des données familiales. La variable aléatoire traduit la variabilité liée à chaque famille. Par exemple, lors des analyses des données familiales de l'étude F5L-Pedigrees, l'effet familial est pris en compte par un effet aléatoire pour la variable représentant la famille.

$$E(Y) = \alpha + \beta_1 * X_{age} + \beta_2 * X_{sexe} + \gamma_3 * X_{famille} + ... + \beta_z * X_{CpG_n} + \epsilon$$

La variable γ_3 suit une loi normale de moyenne 0 et variance σ^2, également estimée par la régression.

D'autres modèles peuvent également être utilisés pour analyser des données de méthylation, les modèles bêta et les modèles quantile. Ils seront abordés dans la partie présentant les perspectives (page 168).

L'application des modèles linéaires/logistiques décrits ci-dessus pour tester l'association entre un site CpG et un phénotype va donner lieu à l'estimation d'un paramètre de régression associé à ce site CpG, de l'écart type de ce paramètre et d'une "p-value". Ces éléments permettent alors d'apprécier si le site CpG considéré est significativement associé, ou pas, à la variabilité du trait étudié. Les résultats des tests d'association entre chaque site CpG et le trait étudié peuvent ensuite être décrits via des graphiques de type Quantile-Quantile (ou Q-Q plot) et "Manhattan" (*Manhattan plot*).

1.4 Diagramme Quantile-Quantile

Le diagramme Quantile-Quantile a pour but de déterminer visuellement si l'on a obtenu plus de résultats significatifs que ne le voudrait le hasard, en d'autres termes si le nombre de faux positifs n'est pas trop élevé (figure 4.1). Pour cela, on compare la distribution des $-log_{10}(Pvalues)$ observées

(en ordonnée sur le graphique) à celle attendue sous l'hypothèse nulle d'absence d'associations entre le phénotype et les niveaux de méthylation. Sous cette hypothèse nulle, les p-values attendues suivent une loi uniforme sur l'intervalle de 0 à 1. On s'attend donc à ce que la distribution des p-values observées ne s'éloigne pas trop de celle des p-values théoriques, qui sous l'hypothèse nulle suit une loi uniforme.

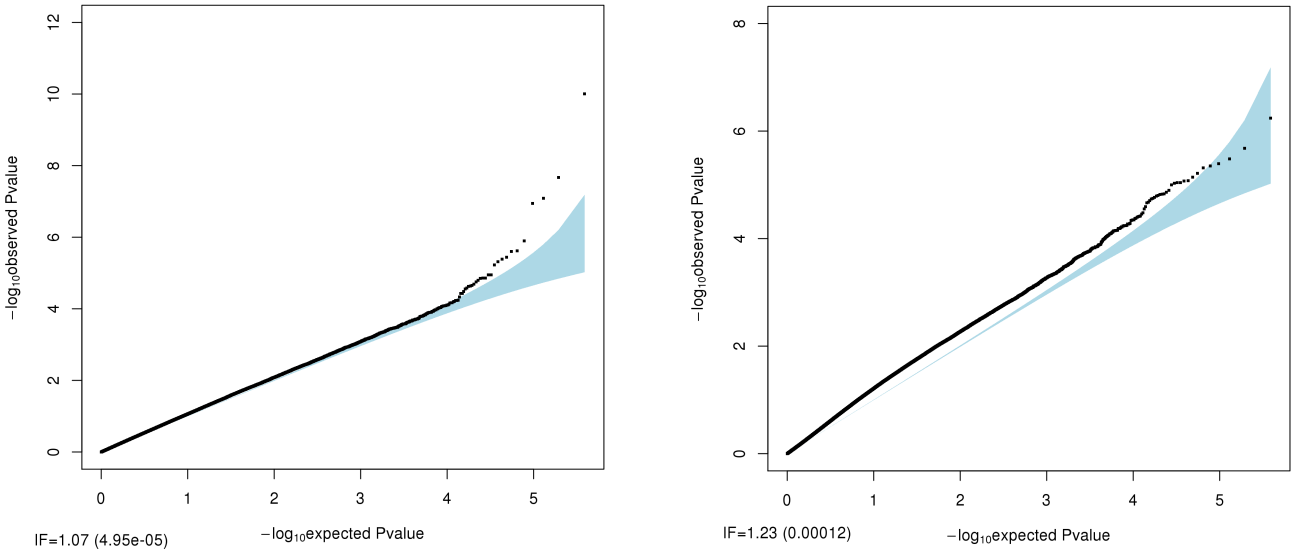

FIGURE 4.1 – Exemple de QQ-plot avec une distribution correcte à gauche (coefficient directeur de la droite de 1,07) et avec une inflation à droite (coefficient directeur de la droite de 1,23) d'une étude MWAS sur le taux de facteur de von Willebrand dans les études MARTHA (gauche) et F5L-Pedigrees (droite).

Il est possible d'obtenir facilement la distribution théorique des p-values des N sites CpG en calculant pour la i^{me} p-value, la p-value théorique $Pval_i$ correspondante :

$$Pval_i = \frac{i}{N+1}$$

L'intervalle de confiance est calculé à partir du fait que les statistiques d'ordre d'une loi uniforme standard suivent une distribution bêta. Pour chaque valeur de p-value, les bornes supérieure et inférieure de l'IC à 95% correspondent respectivement aux quantiles 0,975 et 0,025 de la distribution bêta avec pour paramètres α et β. Où α est le rang de la p-value parmi les p-values ordonnées selon un ordre croissant et β est le rang de la p-value parmi les p-values ordonnées selon un ordre décroissant.

Les p-values observées peuvent également être représentées par un graphique de type Manhattan.

1.5 Graphique Manhattan

Le graphique de type Manhattan (*Manhattan plot*) est un graphique en nuage de points représentant les p-values ($-log_{10}$) obtenus lors d'analyses d'association à grande échelle (comme les GWAS ou MWAS) où chaque point représente un site CpG avec en abscisse la position génomique du site CpG et en ordonnée le $-log_{10}(Pvalue)$ (figure 4.2). Il est généralement commun de tracer une droite horizontale ayant comme ordonnée le seuil de significativité utilisé dans l'étude, voir la partie suivante sur la correction des tests multiples (page 96). Toute association correspondant à un point situé au-dessus de cette ligne de significativité est généralement sélectionnée pour être validée dans un (ou plusieurs) échantillons indépendants pour vérifier sa robustesse, et exclure tout risque de faux positifs associé au seuil statistique utilisé.

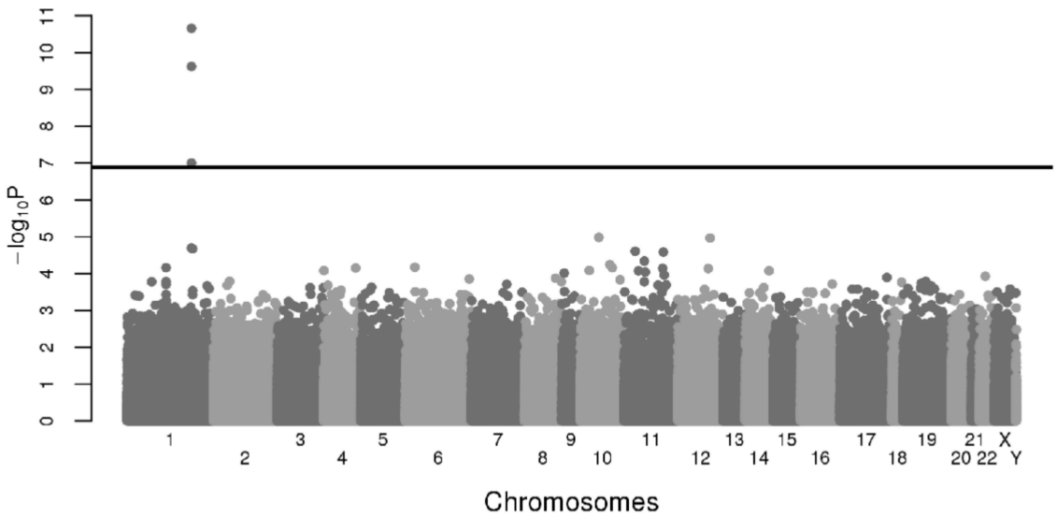

FIGURE 4.2 – Exemple de graphique Manhattan d'une étude MWAS sur la mutation du facteur V de Leiden dans l'étude MARTHA ayant un locus significatif (Chromosome 1, *SLC19A2*). Extrait de AÏSSI *et al.* 2014, faisant partie de ce travail de thèse (voir la partie sur la recherche de facteur épigénétique pouvant expliquer la pénétrance incomplète du Facteur V de Leiden, p. 142).

2 Correction des tests multiples

L'utilisation d'une statistique de test pour évaluer une hypothèse spécifique peut générer deux types d'erreurs : les erreurs de type 1 qui correspondent aux taux de faux positifs (noté α) et les erreurs de type 2 qui correspondent aux taux de faux négatifs (noté β). Le faux positif correspond au cas où l'hypothèse nulle testée serait vraie mais la statistique de test la considérerait comme fausse en la rejetant. Le faux négatif correspond à la situation inverse, c'est à dire que l'hypothèse nulle testée

serait fausse alors que la statistique de test ne la rejette pas, c'est à dire la considérerait comme vraie. Il est donc très important de contrôler ces erreurs pour éviter de faire des conclusions erronées. Pour cela l'approche employée est de réduire au maximum β, donc par conséquent d'augmenter la puissance du test $(1 - \beta)$ tout en gardant α relativement bas. Le seuil communément utilisé est $\alpha = 0,05$, il s'agit d'un seuil fixé arbitrairement. Le fait d'augmenter le nombre de tests effectués va entraîner une augmentation proportionnelle avec le nombre de tests (m) de la probabilité que le test sera déclaré à tort significatif :

$$P(Erreur\ de\ type\ 1) = 1 - (1 - \alpha)^m$$

Les études d'association à l'échelle du génome (GWAS) ou plus précisément ici du méthylome (MWAS) consistent en une répétition du test statistique par le nombre de sites CpG contenus sur la puce, donc plusieurs centaines de milliers de tests. Ce type d'étude nécessite donc d'adopter un seuil statistique très bas pour déclarer une association comme significative pour éviter tous faux positifs et donc toutes conclusions erronées. Différentes méthodes (Dudoit *et al.* 2003 ; Pollard *et al.* 2005) existent pour prendre en compte ce problème de tests multiples. Les 2 principales méthodes utilisées se basent sur le taux d'erreur global (*Family-Wise Error Rate* ou FWER) ou sur le taux de faux positifs (*False Discovery Rate* ou FDR).

2.1 Le taux d'erreur global (FWER)

L'objectif des méthodes se basant sur le taux d'erreur global est de contrôler l'erreur de type I globale sur l'ensemble des tests réalisés. La méthode de Bonferroni est la méthode la plus simple et la plus utilisée. Elle se base sur la probabilité d'obtenir au moins un faux positif dans le cas où tous les tests conduisent au non rejet des hypothèses nulles (c'est à dire qu'ils conduisent à l'absence d'association entre les variables).

La correction de Bonferroni consiste à diviser le seuil de significativité (noté α) par le nombre de tests indépendants effectués (noté N) pour obtenir un nouveau seuil de significativité (noté α').

$$\alpha' = \frac{\alpha}{N}$$

Cette méthode est optimale pour éviter les faux positifs lorsque l'on répète un nombre connu de tests. Dans les cas des études d'associations au niveau du génome, que ce soit pour l'étude des polymorphismes ou de la méthylation des sites CpG, l'indépendance totale entre tous les tests ne peut

être garantie. Dans cette situation, cette méthode devient trop stricte et entraîne une augmentation du nombre de faux négatifs, c'est-à-dire une perte de puissance. Pour remédier à ce problème et ainsi garantir un taux de négatifs relativement bas, une autre famille de correction entre en jeu : le *False Discovery Rate* pour le taux de faux positifs.

2.2 Le taux de faux positifs (FDR)

Une méthode alternative permettant de contrôler le taux de faux positifs (*False Discovery Rate* ou *FDR*) seulement parmi les résultats positifs est proposée par Benjamini et Hochberg (Benjamini & Hochberg 1995 ; Storey & Tibshirani 2003). La correction proposée est d'estimer le taux de faux positifs en calculant le ratio entre le nombre attendu de tests significatifs par erreur et le nombre déclaré de tests significatifs :

$$FDR = \frac{N * \alpha}{k}$$

— N : Nombre de tests effectués

— α : Risque d'erreur sous $H0$, généralement de 0,05

— k : Nombre de tests déclarés significatifs

Cette méthode permet pour chaque test de définir un seuil de significativité α moins strict (donc un α plus élevé) tant que le taux de faux positifs reste inférieur au seuil α_{FDR} que l'on a choisi (généralement 5%). De plus, elle permet d'estimer le taux de faux positifs (appelé q-value) parmi tous les tests ayant des p-values plus petites que celle du test. L'interprétation de la q-value se fait de la même façon que pour la p-value, c'est-à-dire que l'on pourra déclarer un test significatif lorsque la valeur de la q-value sera plus petite que la valeur de α_{FDR}. Le seuil de significativité calculé via le FDR sera proche de celui calculé par la méthode de Bonferroni lorsque le nombre de tests sous H1 sera faible. Lorsque le nombre de tests sous H1 sera élevé, alors le seuil de significativité sera moins strict que celui de Bonferroni.

Une fois les premières analyses terminées, il est nécessaire d'effectuer une étape de réplication.

3 Réplication des résultats de la MWAS

L'étape de réplication des résultats est très importante. Elle permet de confirmer les résultats trouvés dans une étude. En effet, l'association statistique trouvée lors d'une étude d'association méthylome entier peut être due au hasard en faisant partie des 5% d'erreurs autorisées. La réplication

permet de limiter ce risque en effectuant la même analyse dans une étude indépendante. Le fait de retrouver un effet estimé similaire à l'étude précédente et une p-value significative permet de renforcer la confiance dans les résultats trouvés dans la première étude.

4 Méta-Analyses

Le principe de la méta-analyse est de combiner les résultats obtenus dans différentes études pour augmenter la puissance statistique et ainsi trouver des associations ayant des effets plus modestes. Cela est d'autant plus utile lorsqu'il est impossible d'effectuer une analyse combinée des données de différentes études pour différentes raisons, par exemple lorsque :

— Les variables d'ajustements sont différentes. Dans l'étude MARTHA, un analyseur d'hématologie cellulaire a compté le nombre de leucocytes suivant : lymphocyte, monocyte, basophile, éosinophile et neutrophile. Alors que pour l'étude F5L-Pedigrees, l'estimation statistique a permis d'avoir les proportions des leucocytes suivant : lymphocyte T CD4+, lymphocyte T CD8+, lymphocyte NKT, lymphocyte B, monocyte et granulocyte. Dans MARTHA, les proportions des différents granulocytes sont plus détaillées (basophile, éosinophile et neutrophile), alors que dans F5L-Pedigrees, il s'agit des proportions des différents lymphocytes qui le sont plus (T CD4+, T CD8+, NKT et B).

— La conception des études diffère et impose des modèles statistiques différents. Dans l'étude F5L-Pedigrees, les individus ne sont pas indépendants car répartis en différentes familles. Il est donc nécessaire de prendre en compte la corrélation génétique familiale via par exemple un modèle à effets mixtes. Alors que dans l'étude MARTHA, les individus étant indépendants, il n'est pas nécessaire d'utiliser un modèle à effets mixtes mais un simple modèle linéaire suffit.

Différentes méthodes existent pour combiner les résultats, soit en combinant les p-values soit en combinant les effets estimés.

4.1 Combiner les p-values

La première méthode, la plus simple, proposée par Fisher (Fisher 1932), permet d'obtenir une nouvelle statistique (notée X) en combinant les N (pour N études à combiner) p-values (notée P) grâce à la formule suivante :

$$X = -2\sum_{i=1}^{N} ln(P_i)$$

La statistique X suit une distribution χ^2_{2N} ($2 * N$ degrés de liberté), ce qui permet ainsi d'obtenir la nouvelle p-value combinée.

Cette méthode est problématique dans le cas où l'on aurait des effets estimés ayant des sens opposés entre les différentes études. Dans le cas de figure où l'on aurait deux résultats avec des effets estimés opposés et des p-values significatives, le fait de les combiner par cette méthode risque de conduire à une p-value combinée significative. On se retrouverait alors avec des effets estimés opposés entre les deux études et une p-value combinée significative ce qui est incohérent. Pour remédier à ce problème, il est plus judicieux d'utiliser une méthode combinant les effets estimés.

4.2 Combiner les effets estimés

Le second type de méthode pour réaliser des méta-analyses, très largement répandu dans les études génomiques, se base sur la combinaison des coefficients de régression obtenus dans les différentes études. Cette méthode, basée sur l'inverse de la variance pondérée, calcule à partir des effets estimés β_i et de leur écart type SE_i (où $i = 1, 2, ..., N$ et N le nombre d'études à combiner), un effet estimé combiné $\widehat{\beta}$ et un écart type estimé $\widehat{SE}$ ainsi que la p-value correspondante. Soit W_i, le poids accordé à l'étude i et défini ainsi :

$$W_i = \frac{1}{SE_i^2}$$

L'effet estimé combiné $\widehat{\beta}$ est obtenu par la formule :

$$\widehat{\beta} = \frac{\sum_{i=1}^{N} \beta_i W_i}{\sum_{i=1}^{N} W_i}$$

L'écart type estimé $\widehat{SE}$ correspondant à $\widehat{\beta}$ peut se calculer ainsi :

$$\widehat{SE} = \sqrt{\frac{1}{\sum_{i=1}^{N} W_i}}$$

Il est ensuite possible d'obtenir une statistique qui suit une distribution χ^2_1 (1 degré de liberté) en utilisant la formule suivante :

$$X^2 = \frac{\widehat{\beta}^2}{\widehat{SE}^2}$$

Les coefficients obtenus via les formules ci-dessus, appelés aussi estimateurs de méta-analyse à effets fixes, ne sont valides qu'en absence d'hétérogénéité dans les coefficients de régression entre les

différentes études, c'est-à-dire lorsque les coefficients de régression obtenus au sein des différentes études peuvent être considérés comme des estimations d'un même paramètre. Dans le cas d'hétérogénéité, il est préférable d'utiliser les estimateurs de méta-analyse à effets aléatoires qui la prennent en compte. L'absence d'hétérogénéité peut être testée par les deux statistiques (Ioannidis *et al.* 2007) que sont : la statistique Q de Cochran-Mantel-Haenzel et la statistique I^2 de Thomas et Higgins. La statistique Q de Cochran-Mantel-Haenzel s'obtient selon la formule :

$$Q = \sum_{i=1}^{N} \frac{(\widehat{\beta} - \beta_i)^2}{SE_i^2}$$

La p-value associée à la statistique Q s'obtient en comparant Q à une loi χ^2_{N-1} ($N-1$ degrés de liberté). Cependant, ce test est de faible puissance lorsque l'on combine peu d'études. Il est préférable d'utiliser la statistique plus robuste proposée par Thomas et Higgins (Higgins & Thompson 2002 ; Huedo-Medina *et al.* 2006). Celle-ci mesure le pourcentage de variation totale entre les études dû à l'hétérogénéité au-delà de la chance :

$$I^2 = \frac{Q - (N-1)}{Q} * 100$$

La statistique I^2 prend une valeur comprise entre 0 et 100%, une valeur supérieure à 50% indique une forte hétérogénéité.

En cas de présence d'hétérogénéité entre les études, il est nécessaire d'utiliser une méta-analyse à effets aléatoires pour corriger la déflation dans la variance des effets fixes estimés. Elle consiste en l'ajout dans les formules précédentes d'un τ^2, ce qui permet de gonfler la variance des effets estimés dans chaque étude. La correction τ^2 est définie ainsi :

$$\tau^2 = \frac{Q - (N-1)}{\sum_{i=1}^{N} W_i - \frac{\sum_{i=1}^{N} W_i^2}{\sum_{i=1}^{N} W_i}}$$

Elle permet de calculer un nouveau poids W_i^* pour chaque étude :

$$W_i^* = \frac{1}{\tau^2 + SE_i^2}$$

Ainsi que l'effet estimé combiné $\widehat{\beta}^*$ et l'écart type estimé $\widehat{SE}^*$

$$\widehat{\beta}^* = \frac{\sum_{i=1}^{N} \beta_i W_i^*}{\sum_{i=1}^{N} W_i^*}$$

$$\widehat{SE}^* = \sqrt{\frac{1}{\sum_{i=1}^{N} W_i^*}}$$

Il est également possible d'obtenir une statistique qui suit une distribution χ_1^2 (1 degré de liberté) en utilisant la formule suivante :

$$X^{*2} = \frac{\widehat{\beta}^{*2}}{\widehat{SE}^{*2}}$$

Les méta-analyses présentées dans ce travail de thèse ont été réalisées avec le logiciel GWAMA (pour Genome-Wide Association Meta Analysis) (Mägi & A. P. Morris 2010).

Chapitre 5

Identification de marques de méthylation dans l'ADN sanguin associées à des biomarqueurs du risque cardiovasculaire

1 Association entre la méthylation de l'ADN au locus *HIF3A* et l'Indice de Masse Corporelle

Un de mes premiers travaux de thèse a été la participation à une étude internationale du méthylome en relation avec l'Indice de Masse Corporel (IMC).

L'obésité et ses comorbidités associées sont un problème de santé publique majeur et en pleine expansion (Guh *et al.* 2009 ; Huang *et al.* 2015 ; Ogden *et al.* 2014). L'IMC, qui est la mesure la plus répandue de l'obésité, est un phénotype complexe déterminé par l'interaction entre le style de vie (par exemple l'activité physique), des facteurs environnementaux (prises alimentaires et disponibilité de la nourriture) et des facteurs génétiques. Ces dernières années, les recherches ont montré qu'une trentaine de polymorphismes étaient associés à l'IMC, et qu'ensemble, ils expliquaient environ 1 à 5% de la variabilité interindividuelle de l'IMC. Bien que des associations entre plusieurs variants génétiques et l'IMC aient été identifiées, il y a actuellement beaucoup moins de connaissances sur les modifications épigénétiques associées à l'IMC. En collaboration avec l'équipe du Pr. Nilesh Samani (University of Leicester, UK), nous avons entrepris une étude d'association méthylome entier à partir du sang périphérique pour identifier des profils de méthylation qui serait associés à l'IMC (Dick *et al.*

2014). L'étude a porté sur les données de la puce HM450k générées dans différentes cohortes :

- Cardiogenics : échantillons sanguins provenant de 239 patients ayant eu un infarctus du myocarde et 220 sujets sains (Garnier *et al.* 2013) ;
- MARTHA : échantillons sanguins provenant de 339 patients ayant eu une thrombose veineuse ;
- KORA : échantillons sanguins provenant de 1789 sujets de la population générale d'origine allemande (Wichmann *et al.* 2005) ;
- MuTHER : 635 échantillons de tissus adipeux et de 395 échantillons de peau provenant de sujets jumeaux de la population générale d'origine du Royaume-Uni (Nica *et al.* 2011).

Une étude d'association méthylome entier a d'abord été réalisée sur la cohorte de découverte Cardiogenics par l'équipe du Pr. Samani pour identifier des sites CpG dont le niveau de méthylation était fortement associé à l'IMC. Les niveaux de méthylation de 5 sites CpG ont été ainsi trouvés significativement, pour un FDR de 5%, associé à l'IMC ($p < 5,36 * 10^{-7}$). Un site CpG se situait sur le gène *CLUH*, un autre sur le gène *KLF13* et les trois derniers sur le gène *HIF3A*. J'ai eu en charge la première étape de réplication de ces 5 associations dans l'étude MARTHA dont on m'avait confié l'analyse statistique. Seuls les trois sites CpG du gène *HIF3A* ont été retrouvés significativement associés, après correction de Bonferroni, à l'IMC ($p < 5,09 * 10^{-3}$). Suite aux résultats positifs des trois sites CpG du gène *HIF3A* dans l'étude MARTHA, il a été décidé de valider de nouveau ces associations dans les cohortes de réplication secondaires KORA et MuTHER. Le taux de méthylation de ces 3 sites CpG a également été retrouvé significativement associé à l'IMC dans la cohorte KORA ($2,18 * 10^{-3} < p < 0,011$) et dans les échantillons de tissus adipeux de la cohorte MuTHER ($p < 1,72 * 10^{-5}$). Aucune association n'a en revanche été retrouvée dans les échantillons d'ADN issus de la peau dans la cohorte MuTHER.

Ces résultats montrent que l'augmentation de l'IMC chez les adultes d'origine européenne est associée avec une augmentation de la méthylation du gène *HIF3A* dans les cellules sanguines et dans le tissu adipeux. Par exemple, une augmentation de 0,1 du niveau de méthylation du site CpG cg22891070 est associée à une augmentation de l'IMC de 3,6% dans la cohorte Cardiogenics, de 2,7% dans l'étude MARTHA et de 0.8% dans la cohorte KORA. De plus, cette augmentation était retrouvée plus importante, 6,2%, avec les niveaux de méthylation mesurés dans le tissu adipeux des sujets de la cohorte MuTHER. Nous avons ensuite regardé si les polymorphismes du locus *HIF3A* qui influençaient les niveaux de méthylation des sites CpG identifiés pouvaient être associés à l'IMC dans les cohortes utilisées mais également dans les cohortes du consortium GIANT incluant plus de 130 000 individus.

Aucune association convaincante ($p > 0,15$) n'a été observée suggérant que l'IMC influence les niveaux de méthylation du locus *HIF3A* et non l'inverse. Dans le cas contraire, nous aurions dû observer une association, même modeste, entre les SNPs influençant les taux de méthylation et l'IMC. L'influence de l'IMC sur la méthylation d'*HIF3A* n'est probablement pas directe mais peut être expliquée par la présence de facteurs de confusion influençant ces deux variables. L'obésité un facteur de risque du syndrome d'apnées obstructives du sommeil (SAOS) qui lui entraîne une hypoxie (faible concentration d'oxygène) intermittente. Or, le gène *HIF3A* code la protéine de la sous-unité alpha 3 du facteur de transcription induit par l'hypoxie (HIF). Ce dernier régule une grande variété de réponses cellulaires et physiologiques lors d'hypoxie. Par conséquent, l'activation chronique des gènes *HIF* en réponse au SAOS pourrait entraîner un changement de la méthylation des gènes *HIF*. L'ensemble de ces travaux suggère que des perturbations de la voie métabolique du gène *HIF3A* pourraient avoir un rôle important dans la réponse biologique à l'augmentation du poids. Une meilleure compréhension de ces mécanismes pourrait identifier de nouvelles cibles thérapeutiques pour contrer les effets de l'obésité et de ces comorbidités. Il est important de souligner qu'aucune association entre le locus *HIF3A* et l'IMC n'a été trouvée par l'approche GWAS sur plus de 100 000 sujets (Speliotes *et al.* 2010) et plus récemment sur une GWAS d'environ 300 000 sujets (Locke *et al.* 2015). Et ce alors que le travail auquel j'ai participé et ceux réalisés depuis (Demerath *et al.* 2015 ; Pan *et al.* 2015 ; Rönn *et al.* 2015) montrent que l'approche MWAS est complémentaire à l'étude des polymorphismes génétiques pour trouver de nouveaux mécanismes physiopathologiques et que "tout" ne réside pas dans l'étude de la variabilité génétique. L'ensemble des résultats issus de ce travail est décrit dans la publication ci-après.

ARTICLES

DNA methylation and body-mass index: a genome-wide analysis

Katherine J Dick[1,3], Christopher P Nelson[1,3], Loukia Tsaprouni[4,5], Johanna K Sandling[4,6], Dylan Aïssi[7,8,9], Simone Wahl[10,11,12], Eshwar Meduri[4], Pierre-Emmanuel Morange[14], France Gagnon[15], Harald Grallert[10,11,12], Melanie Waldenberger[11,12], Annette Peters[11,12,16], Jeanette Erdmann[17,18], Christian Hengstenberg[16,19], Francois Cambien[7,8,9], Alison H Goodall[1,3], Willem H Ouwehand[4,20,21], Heribert Schunkert[16,19], John R Thompson[2], Tim D Spector[22], Christian Gieger[13], David-Alexandre Trégouët[7,8,9], Panos Deloukas[4,23,24] and Nilesh J Samani[1,3*]

*Correspondence: njs@le.ac.uk
[1] Department of Cardiovascular Sciences, University of Leicester, Leicester, UK
[3] National Institute for Health Research Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, UK
Full list of author information is available at the end of the article

NOTE:
This file is an Author's Post-print version using the BioMed Central TeX template.
For the Publisher's Version/PDF, please see :
The Lancet, 2014 Jun 7;383(9933):1990-8.
DOI: 10.1016/S0140-6736(13)62674-4.

Summary

Background

Obesity is a major health problem that is determined by interactions between lifestyle and environmental and genetic factors. Although associations between several genetic variants and body-mass index (BMI) have been identified, little is known about epigenetic changes related to BMI. We undertook a genome-wide analysis of methylation at CpG sites in relation to BMI.

Methods

479 individuals of European origin recruited by the Cardiogenics Consortium formed our discovery cohort. We typed their whole-blood DNA with the Infinium HumanMethylation450 array. After quality control, methylation levels were tested for association with BMI. Methylation sites showing an association with BMI at a false discovery rate q value of 0.05 or less were taken forward for replication in a cohort of 339 unrelated white patients of northern European origin from the MARTHA cohort. Sites that remained significant in this primary replication cohort were tested in a second replication cohort of 1789 white patients of European origin from the KORA cohort. We examined whether methylation levels at identified sites also showed an association with BMI in DNA from adipose tissue (n=635) and skin (n=395) obtained from white female individuals participating in the MuTHER study. Finally, we examined the association of methylation at BMI-associated sites with genetic variants and with gene expression.

Findings

20 individuals from the discovery cohort were excluded from analyses after quality-control checks, leaving 459 participants. After adjustment for covariates, we identified an association (q value ≤ 0.05) between methylation at five probes across three different genes and BMI. The associations with three of these probes—cg22891070, cg27146050, and cg16672562, all of which are in intron 1 of *HIF3A*—were confirmed in both the primary and second replication cohorts. For every 0.1 increase in methylation β value at cg22891070, BMI was 3.6% (95% CI 2.4–4.9) higher in the discovery cohort, 2.7% (1.2–4.2) higher in the primary replication cohort, and 0.8% (0.2–1.4) higher in the second replication cohort. For the MuTHER cohort, methylation at cg22891070 was associated with BMI in adipose tissue ($p = 1.72 * 10^{-5}$) but not in skin ($p = 0.882$). We observed a significant inverse correlation ($p = 0.005$) between methylation at cg22891070 and expression of one *HIF3A* gene-expression probe in adipose tissue. Two single nucleotide polymorphisms—rs8102595 and rs3826795—had independent associations with methylation at cg22891070 in all cohorts. However, these single nucleotide polymorphisms were not significantly associated with BMI.

Interpretation

Increased BMI in adults of European origin is associated with increased methylation at the *HIF3A* locus in blood cells and in adipose tissue. Our findings suggest that perturbation of hypoxia inducible transcription factor pathways could have an important role in the response to increased weight in people.

Funding

The European Commission, National Institute for Health Research, British Heart Foundation, and Wellcome Trust.

Introduction

Obesity and its associated comorbidities constitute a major and growing health problem worldwide [1]. Therefore, understanding the mechanisms that affect body-mass index (BMI)—the most widely used measure of obesity—and any downstream effects is an important health priority. BMI is a complex phenotype determined by lifestyle (eg, physical activity), environmental factors (food availability and intake), and genetic factors [2]. In the past few years, a major effort to identify genetic determinants of BMI through genome-wide association studies has shown that more than 30 single nucleotide polymorphisms (SNPs) are associated with BMI, which together explain about 1.5% of interindividual variation in BMI [3].

DNA methylation is the reversible and heritable attachment of a methyl group to a nucleotide. The most common form of DNA methylation occurs at the 5' carbon of cytosine in CpG dinucleotides, creating 5-methylcytosine [4]. CpG dinucleotides are often located in CpG islands (clusters of CpG sites) within the promoter region or first exon of genes, or upstream from genes within CpG island shores (DNA regions within 2 Kb of CpG islands) or shelves (within 2 Kb of shores) [4]. DNA methylation plays a part in transcriptional regulation of genes and miRNAs [5], control of alternative promoter usage [6, 7], and alternative splicing [6].

Both genetic and environmental factors can affect the extent of DNA methylation [8, 9]. In view of the range of potential downstream functional outcomes of this epigenetic change, an effect on DNA methylation could integrate the impact of both genetic and environmental factors on a phenotype [10]. Alternatively, epigenetic changes caused by a phenotype can mediate its downstream effects by changing gene expression [10].

Unlike genome-wide association studies of genetic variants, progress in systematic analysis of DNA methylation has hitherto been hampered by an absence of analogous platforms to study epigenetic phenomena. However, the newly developed Infinium HumanMethylation450 array (Illumina, San Diego, CA, USA) assays about 485000 methylation sites spanning 99% of genes in the Reference Sequence database, with an average of 17 CpG sites per gene region. The array has been validated and consistently detects CpG methylation changes [11]. We used this array for a large-scale analysis of methylation patterns in whole-blood DNA in relation to BMI.

Methods

Participants

479 white individuals who had been recruited by the Cardiogenics Consortium [12] formed our discovery cohort. They either had a history of myocardial infarction (n=241; recruited from four centres: Leicester, UK; Lübeck, Germany; Regensburg, Germany; and Paris, France) or were healthy blood donors (n=238; recruited in Cambridge, UK). Genome-wide SNP genotypes had been previously obtained for all participants with the Human Quad Custom 670 array (Illumina, San Diego, CA, USA) and genome-wide gene expression data obtained for monocytes and derived macrophages with the HumanRef-8 v3 Beadchip array (Illumina, San Diego, CA, USA) [13].

For our primary replication cohort, we used data for 339 unrelated white patients of French origin who had venous thrombosis recruited into the MARseille THrombosis Association (MARTHA) cohort [14]. These patients had been genotyped with the Human 610/660W-Quad arrays (Illumina, San Diego, CA, USA) [14].

We analysed methylation sites that showed a significant association in the primary replication cohort in a second replication cohort of 1789 white participants from Germany who had been recruited for the KORA (Cooperative Health Research in the Region of Augsburg) F4 survey [15]. Genome-wide genotyping was done for KORA F4 with the Affymetrix 6.0 GeneChip array (Santa Clara, CA, USA).

To investigate whether the association between methylation at HIF3A sites and BMI that we observed in blood DNA would also be seen in other tissues, we analysed data for white female individuals from the UK obtained as part of the Multiple Tissue Human Expression Resource (MuTHER) study [16]. HumanMethylation450 arrays had been done for 635 subcutaneous adipose tissue biopsies and for 395 skin biopsies. The adipose tissue samples came from 249 twin pairs (93 monozygotic and 156 dizygotic twins) and 137 singletons. Skin samples came from 108 of the 249 twin pairs (44 monozygotic and 64 dizygotic) and 179 singletons. The collection and processing of the biopsy samples in the MuTHER study have been described previously [17]. In addition to the methylation arrays, genome-wide genotype data

(obtained with a combination of HumanHap300, HumanHap610, and 1M-Duo and 1.2M-Duo Illumina arrays; Illumina, San Diego, CA, USA) and genome-wide expression profiles in adipose tissue (obtained with the IlluminaHT-12 v3 array; San Diego, CA, USA) were available for the MuTHER participants [17]. All individuals provided written informed consent to participate in the primary studies and to allow DNA analysis of their samples.

Procedures

Details of the methylation assay done for the discovery cohort and the quality checks that were undertaken are given in the appendix (p 2). Methylation is described as a β value, which is a continuous variable ranging between 0 (no methylation) and 1 (full methylation). In any one sample, a probe with a detection p value (a measure of an individual probe's performance) of more than 0.05 was assigned missing status. If a probe was missing in more than 5% of samples, we excluded it from all samples. We excluded 830 probes on this basis. To avoid spurious associations, we also excluded probes containing genomic sites where variation is already known according to the HumanMethlyation450 annotation files or the InfiniumHD Methylation SNP list that had a minor allele frequency of more than 1%, leaving 351699 probes. Before analysis, methylation values were corrected for background values and then normalised with SWAN [18] in the R Package minfi. We used the array annotations provided by Illumina (version 1.1) to assign probes to their corresponding genes.

We used the same Illumina HumanMethylation450 array in the replication cohorts and in the MuTHER samples, following similar experimental procedures. We did post-array processing in a similar way for all studies and normalised methylation values before analysis with SWAN [18] for the two blood replication cohorts and by quantile normalisation [19] for the MuTHER study samples.

Statistical analysis

BMI was not normally distributed in the discovery cohort and therefore was transformed on the log scale. Regression analysis of log-transformed BMI with methylation level at each probe was adjusted for age, sex, smoking status, methylation array batch, and centre. Adjustment for centre also adjusted for whether patients had had myocardial infarction. Chip assignment was not associated with BMI and was therefore not included in the model. For models in which the dependent variable (BMI in this case) has been log transformed, the β coefficients from the regression analysis can be interpreted as the change in the dependent variable by 100*(coefficient) for an increase in one unit in the independent variable. Therefore, we present β coefficients as percentage change. A correction for genomic control ($\lambda = 1.092$) was applied (appendix p 11). We estimated q values for false discovery rates [20] and associations with a false discovery rate q value of 0.05 or less were taken forward for replication.

We did sequential replication for the MARTHA and KORA cohorts with linear regression analysis of log-transformed BMI adjusted for age, sex, smoking status, and array batch. We assessed significance after Bonferroni correction.

In the MuTHER cohort, to account for family structure, we fitted a linear mixed effects model for log-transformed BMI with the lme4 package in R. We adjusted the

model for age, array batch, and smoking status (fixed effects), and for family identification number and zygosity (random effects). We used the likelihood ratio test statistic to assess significance and calculated the p value from the χ^2 distribution with one degree of freedom.

We assessed associations between methylation level for sites showing a correlation with BMI and genotypes at adjacent SNPs (within 1 Mb) in the discovery cohort, assuming an additive allele effect. We used a linear mixed effects model with age, sex, smoking status, centre, BMI, and methylation batch array as fixed effects, and methylation chip as a random effect. We applied Bonferroni correction for multiple testing to the results. We analysed significant and independent associations in a similar manner in the replication cohorts and in MuTHER samples (with the addition of family identification number and zygosity as random effects and exclusion of sex). We also used the same model to analyse the association between methylation level or BMI with individual blood cell counts in the discovery cohort. We did power calculations with powerreg in Stata (version 12.1).

Role of the funding source

The sponsors of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. PD and NJS had full access to data for the discovery cohort, D-AT to data for the MARTHA cohort, CG for the KORA cohort, and PD for the MuTHER cohort. NJS had the final responsibility for the decision to submit for publication.

Results

20 individuals from the discovery cohort (two who had had myocardial infarction, 18 healthy blood donors) were excluded from analyses after quality-control checks of the methylation array data (appendix p 2), leaving 459 participants (table 1). As reported by others [21] at a genomic level, methylation at CpG dinucleotides in our discovery cohort had a bimodal distribution, with the most frequent level of methylation occurring at a β value of 0–0.05 with a second, slightly lower peak at 0.90–0.95 (appendix p 9). In a previous study (in which the Illumina HumanMethylation27 Bead Chip, the precursor of the HumanMethylation450 Bead Chip, was used) [22], a robust association between current smoking and methylation at the cg03636183 locus in *F2RL3* had been shown and replicated. As a form of overall validation of our discovery analysis, we examined the association of current or ever smoking with methylation at this site in our dataset. We recorded a similarly highly significant association ($p = 3.8 * 10^{-33}$) between methylation at cg03636183 and smoking, with reduced methylation in smokers (appendix p 10).

The distribution of p values in the discovery cohort from regression of methylation level at each site and BMI is shown in figure 1. The quantile–quantile plot for expected versus observed χ^2 values is shown in the appendix (p 11). Five probes achieved a false discovery rate q value of 0.05 or less, including individual probes in *CLUH* on chromosome 15 and *KLF13* on chromosome 17 (appendix p 3), and three probes in *HIF3A* on chromosome 19 (table 2). We excluded the possibility that these probes showed cross-reactivity for several CpG sites [23].

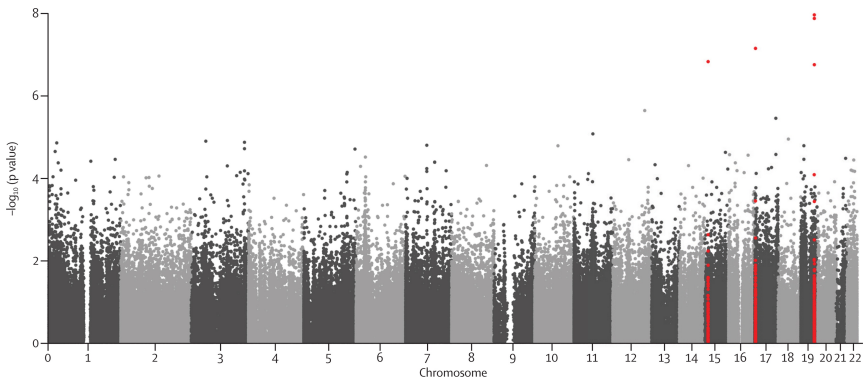

Figure 1 Manhattan plot showing the distribution of p values of the association of methylation probes with body-mass index in the discovery cohort
The red dots indicate probes that fall within *KLF13* (chromosome 15), *CLUH* (chromosome 17), and *HIF3A* (chromosome 19).

Table 1 Characteristics of participants in the studied cohorts

	Discovery cohort (Cardiogenics)		Primary replication cohort (MARTHA; n=339)	Second replication cohort (KORA; n=1789)	MuTHER cohort	
	Individuals who had had myocardial infarction (n=239)	Healthy blood donors (n=220)			Adipose tissue samples (n=635)	Skin samples (n=395) *
Age (years)	55.2 (6.8)	55.2 (6.8)	43.8 (14.2)	60.9 (8.9)	58.8 (9.3)	58.8 (9.3)
Men	202 (85%)	125 (57%)	74 (22%)	871 (49%)	0	0
Body-mass index (kg/m^2)	28.3 (4)	25.9 (3.6)	24.2 (4.4)	28.1 (4.8)	26.7 (4.9)	26.6 (4.7)
Ever smokers	185 (77%)	89 (40%)	145(43%)	1003 (56%)	308 (49%)	187 (47%)
Height (cm)	174.5 (8.7)	172.5 (9.1)	166.6 (7.7)	167.8 (9.2)	161.5 (5.8)	161.5 (6.0)
Weight (kg)	86.5 (15.8)	77.2 (12.5)	67.5 (14.4)	79.4 (15.3)	69.8 (13.8)	69.5 (13.3)
Systolic blood pressure (mm Hg)	130.5 (19.1)	NA	NA	124.8 (18.7)	129.8 (16.2)	129.1 (16.0)
Diastolic blood pressure (mm Hg)	77.8 (10.9)	NA	NA	76.1 (9.9)	78.6 (9.4)	78.6 (9.5)
Diabetic	10 (4%)	NA	6 (2%)	163 (9%)	30 (5%)	16 (4%)
Methylation of cg22891070 †	0.434 (0.110, 0.189-0.910)	0.453 (0.098, 0.211-0.740)	0.473 (0.118, 0.127-0.823)	0.515 (0.131, 0.154-0.906)	0.177 (0.045, 0.076-0.358)	0.272 (0.052, 0.165-0.536)
Methylation of cg27146050 †	0.319 (0.051, 0.144-0.516)	0.328 (0.047, 0.191-0.495)	0.315 (0.042, 0.180-0.458)	0.380 (0.057, 0.179-0.622)	0.163 (0.037, 0.086-0.262)	0.232 (0.029, 0.161-0.368)
Methylation of cg16672562 †	0.389 (0.116, 0.071-0.952)	0.409 (0.101, 0.157-0.745)	0.454 (0.125, 0.107-0.795)	0.438 (0.136, 0.091-0.900)	0.098 (0.039, 0.016-0.237)	0.174 (0.044, 0.064-0.422)

Data are mean (SD), n (%), or mean (SD, range). NA=not available. * From subset of participants who had also provided adipose tissue samples. † β values.

Table 2 Association between methylation at sites in *HIF3A* on chromosome 19 in whole-blood DNA and BMI in the discovery and replication cohorts

	Position	Discovery cohort (Cardiogenics)		Primary replication cohort (MARTHA)		Second replication cohort (KORA)	
		P-value *	Percentage change in BMI (95% CI) †	P-value	Percentage change in BMI (95% CI) †	P-value	Percentage change in BMI (95% CI) †
cg22891070	46801642	$4.00 * 10^{-8}$	3.6% (2.4-4.9)	$3.65 * 10^{-4}$	2.7% (1.2-4.2)	$6.69 * 10^{-3}$	0.8% (0.2-1.4)
cg27146050	46801557	$4.82 * 10^{-8}$	7.8% (5.1-10.4)	$5.09 * 10^{-3}$	6.2% (1.8-10.4)	$2.18 * 10^{-3}$	2.1% (0.7-3.4)
cg16672562	46801672	$5.36 * 10^{-7}$	3.2% (2.0-4.4)	$3.47 * 10^{-3}$	2.1% (0.7-3.5)	0.011	0.7% (0.2-1.3)

The significance threshold after Bonferroni correction for multiple testing in the primary replication cohort is 0.01 and in the second replication cohort is 0.016.
BMI=body-mass index. * λ corrected. † The β coefficients from the association analysis have been converted into percentage change in BMI for every 0.1 unit increase in methylation β value.

We took these five probes forward for analysis in our primary replication cohort (MARTHA). Although methylation level for the probes in *CLUH* and *KLF13* were not associated with BMI in this cohort (appendix p 3), all three *HIF3A* probes were significant after Bonferroni correction for multiple testing (table 2). We further tested the association of these three probes in our second replication cohort (KORA). All three probes were significantly associated with BMI, although the association was weaker than for the other cohorts (table 2).

The three identified *HIF3A* probes (cg22891070, cg27146050, and cg16672562) are neighbouring probes in intron 1 of the gene (figure 2). Methylation levels at cg22891070, cg27146050, and cg16672562 are all highly correlated with each other (R^2 = 0.89–0.95 in the discovery cohort). The three probes are flanked by others that had nominally significant associations with BMI in the discovery cohort (cg05286653: $p = 2.37 * 10^{-4}$; cg12068280: $p = 4.89 * 10^{-3}$) that did not meet our false discovery rate q value threshold of 0.05 or lower. Overall, there are probes for 25 CpG sites in *HIF3A* on the array, and the results for all the probes are shown in the appendix (p 4). Methylation at CpG sites in the other members of the hypoxia inducible transcription factor family (*HIF1A* [13 probes], *EPAS1* [38 probes], and *ARNT* [17 probes]) was not associated with BMI (data not shown).

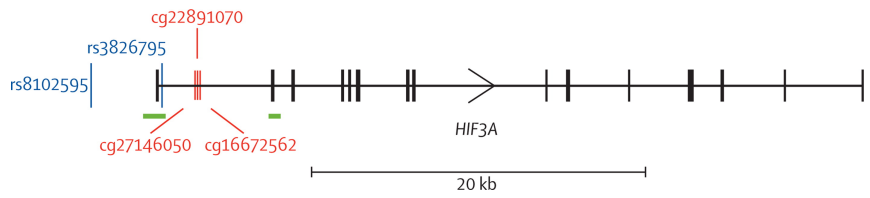

Figure 2 Location of methylation probes associated with body-mass index and SNPs affecting methylation levels of these probes in the HIF3A locus
Vertical black lines represent exons. The arrow indicates direction of transcription. The three methylation sites in intron 1 showing an association with body-mass index are shown in red. The two SNPs showing an association with methylation levels at these probes are shown in blue. The green blocks represent the position of CpG islands in this locus. SNP=single nucleotide polymorphism.

Because the DNA used in our methylation analysis is derived from a mixture of different white blood cell types, methylation in the *HIF3A* probes could vary between different white cell populations, and the correlation with BMI could simply be a result of varying proportions of these cell types in individuals with different BMIs. Therefore, using cg22891070 as an exemplar, we examined the association of methylation level of this probe with the number of each cell type in the discovery cohort using a linear mixed effects model. Additionally, we tested for an association between number of each cell type and BMI. We recorded a weak positive correlation (p = 0.019) between methylation at cg22891070 and lymphocyte count that did not survive correction for multiple testing. We recorded no associations with other cell types (appendix p 5). Furthermore, adjustment for lymphocyte, monocyte, and neutrophil counts did not substantially attenuate the association between methylation at cg22891070 and BMI ($p = 1.04 * 10^{-7}$).

We also examined the association of DNA methylation at *HIF3A* with the two individual components of BMI—height and weight—in the discovery cohort. Methylation at cg22891070 was significantly associated with weight ($p = 5.2 * 10^{-7}$) but

not with height ($p = 0.78$). In exploratory analyses of the population-based KORA cohort, we did not find an association between methylation at cg22890170 and other characteristics associated with BMI, such as physical activity ($p = 0.955$) or type 2 diabetes mellitus ($p = 0.680$).

For the three significant sites in *HIF3A*, overall methylation β value in the discovery cohort ranged from 0.18 to 0.90 for cg22891070, from 0.14 to 0.52 for cg27146050, and from 0.07 to 0.95 for cg16672562 (appendix p 12). β values were similar in the replication cohorts (table 1). The correlation between methylation level at cg22891070 in blood DNA and BMI for the discovery cohort, and the change in methylation level at cg22891070 by quintile of BMI (and vice versa) are shown in the appendix (pp 13–14). Every 0.1 increase in methylation β value for cg22891070 was associated with a 3.6% higher BMI in the discovery cohort (table 2). For a person in the discovery cohort with the mean BMI (27 kg/m^2), this 3.6% increase equates to a 0.98 kg/m^2 higher BMI on average. The increase in BMI was greater in participants who had had myocardial infarction (4.6%, 95% CI 2.9–6.3) than in the blood donors (2.3%, 0.4–4.1). The percentage changes in BMI in the replication cohorts for a 0.1 increase in methylation were smaller than in the discovery cohort (table 2), and in KORA was equivalent to a 0.22 kg/m^2 higher BMI on average.

In the MuTHER cohort, methylation level at the three *HIF3A* sites was strongly associated with BMI in adipose tissue but not in skin (table 3). The range of methylation β values was narrower in both tissues than in blood DNA (table 1). However, it was narrower in adipose tissue than in skin, which means that a reduced range cannot be a reason for why an association was not observed in skin. The direction of the association between methylation in *HIF3A* in adipose tissue and BMI was the same as that in blood, but the percentage change was greater.

Table 3 Association between BMI and methylation at sites in *HIF3A* in adipose tissue and skin DNA in the MuTHER cohort

	Adipose tissue (n=635)		Skin (n=395)	
	P-value	Percentage change in BMI *	P-value	Percentage change in BMI *
cg22891070	$1.72 * 10^{-5}$	6.2 (3.4 to 9.0)	0.882	-0.25 (-3.6 to 3.0)
cg27146050	$9.27 * 10^{-7}$	11.9 (7.2 to 16.7)	0.011	-7.0 (-12.4 to -1.7)
cg16672562	$5.01 * 10^{-6}$	7.9 (4.5 to 11.2)	0.862	-0.36 (-4.3 to 3.5)

Data in parentheses are 95% CIs. BMI=body-mass index. * The β coefficients from the association analysis have been converted into percentage change in BMI for every 0.1 unit increase in methylation β value.

We could analyse whether methylation at the *HIF3A* locus was correlated with *HIF3A* gene expression for the MuTHER adipose dataset, because genome-wide expression profiles were available. We recorded a weak (β value –0.025, SE 0.008) but significant (p=0.005) inverse correlation between methylation at cg22891070 and one (ILMN_1663015) of five *HIF3A* gene-expression probes on the array (appendix p 6). Although we had genome-wide expression data from monocytes and macrophages for the discovery cohort [13], expression of *HIF3A* was below detectable levels in these cells so we could not directly examine whether variation in methylation level at cg22891070 is associated with expression of the gene in blood cells.

Because DNA sequence variation can be associated with methylation level, we looked for an association between SNPs within 1 Mb of cg22891070 and methylation at this probe, using the genome-wide SNP data available for the discovery cohort

(appendix p 15). Two SNPs, rs8102595 and rs3826795, with an R^2 between them of 0.006 ($D' = 1$), had independent associations with methylation at cg22891070 (table 4). rs8102595 had a stronger association than did rs3826795 (table 4). rs8102595 is located 3.8 kb and rs3826795 1.2kb upstream of cg22891070 (figure 2). Associations between these SNPs and methylation at cg22891070 were also highly significant in the replication cohorts (table 4). Furthermore, the same associations were recorded in both adipose tissue and skin in the MuTHER cohort (table 4). Genetic variation in rs8102595 accounted for 6.4% of the variation in methylation at cg22891070 in the blood DNA in the discovery cohort, 9.9% in the MARTHA cohort, and 4.8% in the KORA cohort. This genetic variation also accounted for 14.3% of variation in methylation at cg22891070 in adipose tissue and 21.8% in skin in the MuTHER study.

Table 4 Association between methylation level at cg22891070 and single nucleotide polymorphisms at the *HIF3A* locus

	rs8102595			rs3826795		
	Frequency of effect allele *	β (95% CI)	P-value	Frequency of effect allele †	β (95% CI)	P-value
Discovery (Cardiogenics)	0.10	0.063 (0.042-0.083)	$6.29 * 10^{-9}$	0.81	0.039 (0.023-0.056)	$3.21 * 10^{-6}$
Primary replication cohort (MARTHA)	0.10	0·097 (0·062-0·121)	$1.41 * 10^{-9}$	0.79	0.051 (0.023-0.076)	$2.14 * 10^{-5}$
Second replication cohort (KORA)	0.09	0·073 (0·058-0·086)	$9.18 * 10^{-23}$	0.82	0.048 (0.037-0.059)	$2.26 * 10^{-18}$
MuTHER cohort: adipose tissue	0.10	0·041 (0·033-0·049)	$1.05 * 10^{-21}$	0.81	0.021 (0.014-0.028)	$3.61 * 10^{-9}$
MuTHER cohort: skin	0.10	0·062 (0·052-0·074)	$7.09 * 10^{-25}$	0.82	0.023 (0.013-0.034)	$1.77 * 10^{-5}$

The β values are from an additive model and are a unit change in methylation per copy of the effect allele. * G. † C.

In view of the association between the two SNPs and methylation at cg22891070, we next tested their association with BMI in the discovery and other cohorts, but observed no consistently significant association (appendix p 7). However, the power of these analyses was low (appendix p 7). Therefore, we also tested for associations between these SNPs and indices of body mass in the publicly available GIANT consortium datasets [3]. We found no significant association of either SNP with BMI (rs8102595: n=123 791, p=0.15; rs3826795: n=123 847, p=0.25; appendix p 8).

Discussion

We have identified and replicated a specific association between BMI and methylation of *HIF3A* in whole blood DNA. We recorded the same association in DNA from adipose tissue, which is of high relevance to bodyweight and obesity, implying that it is biologically relevant. Although some preliminary reports are available of whole-blood methylation profiles in relation to indices of body composition and obesity [24, 25, 26, 27], we are the first to have undertaken a large-scale analysis with replication of the principal finding (panel).

Panel: **Research in context**

Systematic review

We searched Medline on Dec 1, 2013, with the terms "BMI & DNA methylation", "obesity & DNA methylation", "BMI & epigenetics" and "obesity & epigenetics". We identified hundreds of reports, many of which were reviews about the potential relevance of epigenetics in obesity. Of original research, some reports focused on methylation of specific genes already known to be associated with body-mass index (BMI) or obesity, such as FTO and POMC. In a few small genome-wide studies [24, 25, 26, 27], the association between methylation and BMI or other indices of obesity has been explored, without definitive findings. One study of overweight or obese adolescents [28] identified five regions that showed differential methylation levels between individuals who had a high and low response to a multidisciplinary weight-loss intervention. Another study showed significant changes in genome-wide methylation pattern in human adipose tissue after a 6-month exercise intervention [29]. Although further validation is necessary, these studies show that DNA methylation can be dynamic and could also affect whether weight changes in response to lifestyle and dietary measures.

Interpretation

Ours is the first large-scale genome-wide analysis of the association between adult BMI and DNA methylation. We have shown that BMI is associated with methylation of HIF3A in blood and adipose tissue. Our findings provide a strong foundation for further exploration of the part played by the epigenome in regulation of BMI and the downstream detrimental effects of increased bodyweight. Understanding of this role could identify novel therapeutic targets to tackle obesity.

HIF3A is a component of the hypoxia inducible transcription factor (HIF), which regulates a wide variety of cellular and physiological responses to reduced oxygen

concentrations by controlling expression of many target genes [30]. It is a heterodimer that is composed of a β subunit (ARNT) and one of three α subunits (HIF1A, EPAS1, and HIF3A). The binding of each α subunit to ARNT targets different sets of downstream genes in a cell-specific manner [30]. In the case of HIF3A, a further layer of complexity is added by the fact that the *HIF3A* locus is subject to much alternate splicing, leading to at least seven variants with differing targets [31]. The induction of target genes by HIF3A binding to ARNT is generally weaker than is that evoked by HIF1A and EPAS1 binding to ARNT [30, 31]. Furthermore, especially in situations in which the amount of ARNT could be limiting, at least some isoforms of HIF3A seem to hinder the response to hypoxia by sequestering ARNT and restricting its binding to HIF1A and EPAS1 [32, 33].

Although the main focus on HIF has been its role in cellular and vascular response to changes in oxygen tension during normal development or pathological processes (eg, cardiovascular disease and cancer [30]), compelling and increasing experimental data suggest that the HIF system also plays a key part in metabolism, energy expenditure, and obesity [34, 35, 36, 37]. Specifically, targeted disruption of either HIF1A or ARNT in adipocytes in transgenic mice is associated with reduced fat formation and protection from obesity and insulin resistance induced by high-fat diets [34]. Similarly, systemic use of an antisense oligonucleotide to HIF1A for 8 weeks in mice with diet-induced obesity substantially suppresses *HIF1A* expression in liver and adipose tissue and is associated with increased energy expenditure and weight loss [35]. In the hypothalamus, HIF signalling (primarily via EPAS1) has a role in glucose sensing and regulation of energy balance and weight by affecting expression of pro-opiomelanocortin [36].

Although HIF3A has not been investigated as thoroughly as the other α subunits in this context, it has been shown to have a role in the cellular response to glucose and insulin, and functions as an accelerator of adipocyte differentiation [38, 39]. Furthermore, siRNA inhibition of *HIF3A* in Hep3B cells significantly downregulates mRNA expression of *ANGPTL4* [31], which could have a role in acquired obesity [40].

The cross-sectional nature of our analysis means that we cannot assign a cause–effect association directly from the association we observed between *HIF3A* methylation and BMI. Previous studies [41, 42] have shown that DNA sequence variation can affect levels of methylation at individual sites (methylation quantitative trait loci). To investigate directionality of the association between *HIF3A* methylation and BMI, we searched for genetic variants that associate with *HIF3A* methylation to establish whether these variants also associate with BMI in turn. We identified significant independent associations between genotypes at two SNPs—rs8102595 and rs3826795, upstream of *HIF3A*—and methylation at one of our identified *HIF3A* probes, cg22891070. However, we identified no association between these variants and BMI in our cohorts or in the large GIANT genome-wide association meta-analysis of BMI which included more than 123000 individuals. Our analysis of GIANT data had more than 95% power to detect an association for both SNPs if one existed (appendix p 8). These findings suggest that the association between increased methylation and higher BMI is not causal. Furthermore, the finding that methylation in *HIF3A* in skin was not associated with BMI, despite

a strong methylation quantitative trait locus for cg22890170 in this tissue, also indicates the absence of causal directionality. Therefore, our findings suggest that increased methylation at the *HIF3A* locus is a result of increased BMI.

An alternative possibility is that the association between methylation at *HIF3A* and BMI is due to a confounding factor which affects both variables. However, we did not observe the association between *HIF3A* methylation and BMI in skin. Furthermore, we did not observe any association with other characteristics associated with BMI, such as physical activity or diabetes.

The mechanism by which increased BMI could lead to rises in *HIF3A* methylation is unknown. Obesity predisposes individuals to obstructive sleep apnoea [43], which is associated with intermittent hypoxia. In turn, hypoxia activates HIF signalling. Therefore, chronic upregulation of HIFs in response to obstructive sleep apnoea could result in secondary changes in methylation of the *HIF* genes. However, the association of methylation level at the *HIF3A* locus showed a linear correlation across the range of BMI levels, and increased methylation was not confined to obese individuals (appendix p 13). Furthermore, the association of BMI with variation in methylation was specific to *HIF3A* and was not noted for *HIF1A* and *EPAS1*.

We identified a significant inverse association between *HIF3A* methylation and *HIF3A* expression in adipose tissue. The association was only recorded with one of five *HIF3A* expression probes on the genome-wide expression array (appendix p 6), suggesting that the effect of methylation could be transcript-specific [31]. In this context, we note that all three CpG sites at the *HIF3A* locus that were associated with BMI are situated within regions of open chromatin as identified by formaldehyde-assisted isolation of regulatory elements (FAIRE) in H1-hESC cells and K562 cells, suggesting that these sites lie in a regulatory region [44]. However, two of the expression probes analysed (ILMN_1663015 and ILMN_1687481) are reported to tag the same *HIF3A* transcript (appendix p 6), and the reason for the discrepant findings for these two probes is unclear. Therefore, further work needs to be done to confirm the effect of methylation on expression and any transcript specificity. However, our finding supports the possibility that even if the association between increased methylation of *HIF3A* and BMI is secondary, an alteration in HIF signalling as a result of obesity-induced *HIF3A* methylation could still have an important role in some of the deleterious downstream effects of the disorder.

Although we recorded significant associations between increased *HIF3A* methylation in blood DNA and increased BMI in three different cohorts, the strength of the association varied substantially across the different cohorts. The gradient of the relation between methylation at *HIF3A* and BMI was four-times steeper in the discovery cohort than in the second population-based replication cohort (KORA), despite a similar distribution of methylation values. Whether this difference represents an element of winner's curse [45] or reflects other variation in the characteristics of the cohorts (including the presence of disease in some) is unclear. Even in the discovery cohort, we noted a difference in the level of association between the individuals who had had myocardial infarction and the healthy blood donors. The strength of the association in the blood donors was similar to that in the MARTHA cohort, which comprised patients with deep vein thrombosis, suggesting that the variation is not entirely related to disease status. Therefore, further studies are needed to identify

factors that affect *HIF3A* methylation and modulate the association between BMI and *HIF3A* methylation in whole-blood DNA. Further work is also necessary to deduce the timing of the variation in methylation at the *HIF3A* locus in relation to BMI and whether it is dynamic or not.

Blood is readily accessible for DNA analyses. By contrast with genetic analyses, a challenge of epigenetic analyses is that circulating leucocytes—the source of DNA in blood—are composed of several different cell subtypes that could each show cell-type specific variation in DNA methylation patterns. To an extent, as we have shown, this variation can be assessed and statistical adjustment done. Perhaps a more fundamental issue for the epigenetics community is whether analysis of blood DNA methylation is worthwhile and can reflect changes in relevant tissues for a phenotype. In this regard, our finding of an association between BMI and specific *HIF3A* methylations sites in both blood and adipose tissue DNA supports the use of whole-blood DNA methylation profiling for identification of relevant epigenetic changes and provides a rationale for other studies of this type.

We used a strict sequential replication design to avoid the penalty of multiple testing for confirmation of the association of probes identified in the discovery cohort. We also started with a fairly small discovery cohort. Therefore, we recognise that we have probably missed associations between methylation of other genes and BMI. Meta-analyses of the datasets used in our study together with other datasets could yield additional insights into epigenetic changes associated with BMI.

In summary, we have reported a novel association of increased BMI in adults of European origin with increased methylation at the *HIF3A* locus in blood cells and in adipose tissue. The finding extends reports linking HIF and obesity in experimental models and provides direct evidence in people that perturbation of HIF signalling could have an important role in mediation of some of the downstream adverse responses to increased BMI.

Contributors

KJD, CPN, PD, and NJS conceived the study. JE, CH, FC, AHG, WHO, HS, and NJS were responsible for recruitment and phenotyping of the discovery (Cardiogenics) cohort. LT and EM generated methylation array data for the discovery cohort. KJD and CPN analysed data for the discovery cohort, supervised by JRT. DA, P-EM, FG, and D-AT provided data from the primary replication cohort (MARTHA) and did analyses. SW, HG, MW, AP, and CG provided data from the second replication cohort (KORA) and did analyses. JKS, TDS, and PD provided data from the MuTHER cohort and did analyses. KJD, CPN, and NJS wrote the report. All authors reviewed the report and provided comments.

Declaration of interests

We declare that we have no competing interests.

Acknowledgements

This work was done as part of the Cardiogenics Project, which is funded by the European Union (LSHM-CT 2006–037593). The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique, and the methylation array typing was funded by the Canadian Institutes of Health Research (grant MOP 86466) and the Heart and Stroke Foundation of Canada (grant T6484). Statistical analyses of the MARTHA datasets were done in the C2BIG computing centre (UPMC, Paris, France), which is funded by the Fondation pour la Recherche Médicale and Région Ile de France. The KORA study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. The MuTHER study was funded by a programme grant from the Wellcome Trust (081917/Z/07/Z), and receives support from the National Institute for Health Research BioResource Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. CPN is funded by the National Institute for Health Research Leicester Cardiovascular Biomedical Research Unit, and this work comes under the portfolio of translational research supported by this unit. DA was supported by a PhD grant from the Région Ile de France (CORDDIM). FG holds a Canada Research Chair. JE, FC, HS, D-AT, and NJS collaborate under a Fondation Leducq Grant (12CVD02). TDS is a European Research Council Senior Investigator and is holder of an ERC Advanced Principal Investigator award.

PD is supported by the Wellcome Trust core grant to the Wellcome Trust Sanger Institute (098051), which funded DNA methylation analysis for MuTHER. NJS holds a chair funded by the British Heart Foundation and is a National Institute for Health Research Senior Investigator. We thank the staff from the genotyping facilities at the Wellcome Trust Sanger Institute for sample preparation, quality control, and typing for the Cardiogenics and MuTHER cohorts.

Author details

[1] Department of Cardiovascular Sciences, University of Leicester, Leicester, UK. [2] Department of Health Sciences, University of Leicester, Leicester, UK. [3] National Institute for Health Research Leicester Cardiovascular Biomedical Research Unit, Glenfield Hospital, Leicester, UK. [4] Human Genetics, Wellcome Trust Sanger Institute, Hinxton, UK. [5] ISPAR Institute, University of Bedforshire, Bedford, UK. [6] Department of Medical Sciences, Molecular Medicine and Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [7] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, F-75013 Paris, France. [8] INSERM, UMR_S 1166, F-75013 Paris, France. [9] ICAN Institute for Cardiometabolism And Nutrition, F-75013 Paris, France. [10] German Center for Diabetes Research, Neuherberg, Germany. [11] Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [12] Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [13] Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [14] INSERM, UMR_S 1062, Aix-Marseille University, Marseille, France. [15] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [16] German Centre for Cardiovascular Research, Munich Heart Alliance, Munich, Germany. [17] Institut für Integrative und Experimentelle Genomik, Universität zu Lübeck, Lübeck, Germany. [18] German Centre for Cardiovascular Research, Hamburg/Kiel/Lübeck, Germany. [19] Deutsches Herzzentrum München, Technische Universität München, Munich, Germany. [20] Department of Haematology, University of Cambridge, Cambridge, UK. [21] National Health Service Blood and Transplant, Cambridge, UK. [22] Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [23] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK. [24] Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia.

References

1. Swinburn, B.A., Sacks, G., Hall, K.D., McPherson, K., Finegood, D.T., Moodie, M.L., Gortmaker, S.L.: The global obesity pandemic: shaped by global drivers and local environments **378**(9793), 804–814. doi:10.1016/S0140-6736(11)60813-1. Accessed 2015-10-15
2. Speakman, J.R., O'Rahilly, S.: Fat: an evolving issue **5**(5), 569–573. doi:10.1242/dmm.010553. Accessed 2015-10-15
3. Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R., Randall, J.C., Vedantam, S., Winkler, T.W., Qi, L., Workalemahu, T., Heid, I.M., Steinthorsdottir, V., Stringham, H.M., Weedon, M.N., Wheeler, E., Wood, A.R., Ferreira, T., Weyant, R.J., Segrè, A.V., Estrada, K., Liang, L., Nemesh, J., Park, J.-H., Gustafsson, S., Kilpeläinen, T.O., Yang, J., Bouatia-Naji, N., Esko, T., Feitosa, M.F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A.V., Welch, R., Zhao, J.H., Aben, K.K., Absher, D.M., Amin, N., Dixon, A.L., Fisher, E., Glazer, N.L., Goddard, M.E., Heard-Costa, N.L., Hoesel, V., Hottenga, J.-J., Johansson, A., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M.F., Myers, R.H., Narisu, N., Perry, J.R.B., Peters, M.J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L.J., Timpson, N.J., Tyrer, J.P., van Wingerden, S., Watanabe, R.M., White, C.C., Wiklund, F., Barlassina, C., Chasman, D.I., Cooper, M.N., Jansson, J.-O., Lawrence, R.W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M.T.S., Almgren, P., Arnold, A.M., Aspelund, T., Atwood, L.D., Balkau, B., Balmforth, A.J., Bennett, A.J., Ben-Shlomo, Y., Bergman, R.N., Bergmann, S., Biebermann, H., Blakemore, A.I.F., Boes, T., Bonnycastle, L.L., Bornstein, S.R., Brown, M.J., Buchanan, T.A., Busonero, F., Campbell, H., Cappuccio, F.P., Cavalcanti-Proença, C., Chen, Y.-D.I., Chen, C.-M., Chines, P.S., Clarke, R., Coin, L., Connell, J., Day, I.N.M., Heijer, M.d., Duan, J., Ebrahim, S., Elliott, P., Elosua, R., Eiriksdottir, G., Erdos, M.R., Eriksson, J.G., Facheris, M.F., Felix, S.B., Fischer-Posovszky, P., Folsom, A.R., Friedrich, N., Freimer, N.B., Fu, M., Gaget, S., Gejman, P.V., Geus, E.J.C., Gieger, C., Gjesing, A.P., Goel, A., Goyette, P., Grallert, H., Gräßler, J., Greenawalt, D.M., Groves, C.J., Gudnason, V., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Hall, A.S., Havulinna, A.S., Hayward, C., Heath, A.C., Hengstenberg, C., Hicks, A.A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K.B., Jarick, I., Jewell, E., John, U., Jørgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L.M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J.W., Kolcic, I., König, I.R., Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaløy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L.J., Lecoeur, C., Lehtimäki, T., Lettre, G., Liu, J., Lokki, M.-L., Lorentzon, M., Luben, R.N., Ludwig, B., Magic, Manunta, P., Marek, D., Marre, M., Martin, N.G., McArdle, W.L., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G.W., Morken, M.A., Morris, A.P., Mulic, R., Ngwa, J.S., Nelis, M., Neville, M.J., Nyholt, D.R., O'Donnell, C.J., O'Rahilly, S., Ong, K.K., Oostra, B., Paré, G., Parker, A.N., Perola, M., Pichler, I., Pietiläinen, K.H., Platou, C.G.P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N.W., Ridderstråle, M., Rief, W., Ruokonen, A., Robertson, N.R., Rzehak, P., Salomaa, V., Sanders, A.R., Sandhu, M.S., Sanna, S., Saramies, J., Savolainen, M.J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D.S., Smit, J.H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A.J., Tammesoo, M.-L., Tardif, J.-C., Teder-Laving, M., Teslovich, T.M., Thompson, J.R., Thomson, B., Tönjes, A., Tuomi, T., van Meurs, J.B.J., van Ommen, G.-J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C.I.G., Voight, B.F., Waite, L.L., Wallaschofski, H., Walters, G.B., Widen, E., Wiegand, S., Wild, S.H., Willemsen, G., Witte, D.R., Witteman, J.C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J.P., Farooqi, I.S., Hebebrand, J., Huikuri, H.V., James, A.L., Kähönen, M., Levinson, D.F., Macciardi, F., Nieminen, M.S., Ohlsson, C., Palmer, L.J., Ridker, P.M., Stumvoll, M., Beckmann, J.S., Boeing, H., Boerwinkle, E., Boomsma, D.I., Caulfield, M.J., Chanock, S.J., Collins, F.S., Cupples, L.A., Smith, G.D., Erdmann, J., Froguel,

P., Grönberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T.B., Hattersley, A.T., Hayes, R.B., Heinrich, J., Hu, F.B., Hveem, K., Illig, T., Jarvelin, M.-R., Kaprio, J., Karpe, F., Khaw, K.-T., Kiemeney, L.A., Krude, H., Laakso, M., Lawlor, D.A., Metspalu, A., Munroe, P.B., Ouwehand, W.H., Pedersen, O., Penninx, B.W., Peters, A., Pramstaller, P.P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N.J., Schwarz, P.E.H., Shuldiner, A.R., Spector, T.D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T.T., Wabitsch, M., Waeber, G., Wareham, N.J., Watkins, H., on behalf of Procardis Consortium, Wilson, J.F., Wright, A.F., Zillikens, M.C., Chatterjee, N., McCarroll, S.A., Purcell, S., Schadt, E.E., Visscher, P.M., Assimes, T.L., Borecki, I.B., Deloukas, P., Fox, C.S., Groop, L.C., Haritunians, T., Hunter, D.J., Kaplan, R.C., Mohlke, K.L., O'Connell, J.R., Peltonen, L., Schlessinger, D., Strachan, D.P., van Duijn, C.M., Wichmann, H.-E., Frayling, T.M., Thorsteinsdottir, U., Abecasis, G.R., Barroso, I., Boehnke, M., Stefansson, K., North, K.E., McCarthy, M.I., Hirschhorn, J.N., Ingelsson, E., Loos, R.J.F.: Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index **42**(11), 937–948. doi:10.1038/ng.686. Accessed 2015-10-15

4. Jones, P.A.: Functions of DNA methylation: islands, start sites, gene bodies and beyond **13**(7), 484–492. doi:10.1038/nrg3230. Accessed 2015-10-15
5. Lopez-Serra, P., Esteller, M.: DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer **31**(13), 1609–1622. doi:10.1038/onc.2011.354. Accessed 2015-10-15
6. Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Sung, K.W.K., Rigoutsos, I., Loring, J., Wei, C.-L.: Dynamic changes in the human methylome during differentiation **20**(3), 320–331. doi:10.1101/gr.101907.109. Accessed 2015-10-15
7. Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., Turecki, G., Delaney, A., Varhol, R., Thiessen, N., Shchors, K., Heine, V.M., Rowitch, D.H., Xing, X., Fiore, C., Schillebeeckx, M., Jones, S.J.M., Haussler, D., Marra, M.A., Hirst, M., Wang, T., Costello, J.F.: Conserved role of intragenic DNA methylation in regulating alternative promoters **466**(7303), 253–257. doi:10.1038/nature09165. Accessed 2015-10-15
8. Lienert, F., Wirbelauer, C., Som, I., Dean, A., Mohn, F., Schübeler, D.: Identification of genetic elements that autonomously determine DNA methylation states **43**(11), 1091–1097. doi:10.1038/ng.946. Accessed 2015-10-15
9. Feil, R., Fraga, M.F.: Epigenetics and the environment: emerging patterns and implications **13**(2), 97–109. doi:10.1038/nrg3142. Accessed 2015-10-15
10. Schadt, E.E.: Molecular networks as sensors and drivers of common human diseases **461**(7261), 218–223. doi:10.1038/nature08454. Accessed 2015-10-15
11. Sandoval, J., Heyn, H., Moran, S., Serra-Musach, J., Pujana, M.A., Bibikova, M., Esteller, M.: Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome **6**(6), 692–702. doi:10.4161/epi.6.6.16196. Accessed 2015-10-15
12. Garnier, S., Truong, V., Brocheton, J., Zeller, T., Rovital, M., Wild, P.S., Ziegler, A., Munzel, T., Tiret, L., Blankenberg, S., Deloukas, P., Erdmann, J., Hengstenberg, C., Samani, N.J., Schunkert, H., Ouwehand, W.H., Goodall, A.H., Cambien, F., Trégouët, D.-A., The Cardiogenics Consortium: Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes **9**(1), 1003240. doi:10.1371/journal.pgen.1003240. Accessed 2015-10-15
13. Heinig, M., Petretto, E., Wallace, C., Bottolo, L., Rotival, M., Lu, H., Li, Y., Sarwar, R., Langley, S.R., Bauerfeind, A., Hummel, O., Lee, Y.-A., Paskas, S., Rintisch, C., Saar, K., Cooper, J., Buchan, R., Gray, E.E., Cyster, J.G., Cardiogenics Consortium, Erdmann, J., Hengstenberg, C., Maouche, S., Ouwehand, W.H., Rice, C.M., Samani, N.J., Schunkert, H., Goodall, A.H., Schulz, H., Roider, H.G., Vingron, M., Blankenberg, S., Münzel, T., Zeller, T., Szymczak, S., Ziegler, A., Tiret, L., Smyth, D.J., Pravenec, M., Aitman, T.J., Cambien, F., Clayton, D., Todd, J.A., Hubner, N., Cook, S.A.: A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk **467**(7314), 460–464. doi:10.1038/nature09386. Accessed 2015-10-15
14. Trégouët, D.-A., Heath, S., Saut, N., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Drouet, L., Zelenika, D., Juhan-Vague, I., Alessi, M.-C., Tiret, L., Lathrop, M., Emmerich, J., Morange, P.-E.: Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach **113**(21), 5298–5303. doi:10.1182/blood-2008-11-190389. Accessed 2015-10-15
15. Wichmann, H.-E., Gieger, C., Illig, T.: KORA-gen - resource for population genetics, controls and a broad spectrum of disease phenotypes **67**, 26–30. doi:10.1055/s-2005-858226. Accessed 2015-10-15
16. Nica, A.C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., Hedman, A.K., Bataille, V., Tzenova Bell, J., Surdulescu, G., Dimas, A.S., Ingle, C., Nestle, F.O., di Meglio, P., Min, J.L., Wilk, A., Hammond, C.J., Hassanali, N., Yang, T.-P., Montgomery, S.B., O'Rahilly, S., Lindgren, C.M., Zondervan, K.T., Soranzo, N., Barroso, I., Durbin, R., Ahmadi, K., Deloukas, P., McCarthy, M.I., Dermitzakis, E.T., Spector, T.D., The MuTHER Consortium: The architecture of gene regulatory variation across multiple human tissues: The MuTHER study **7**(2), 1002003. doi:10.1371/journal.pgen.1002003. Accessed 2015-10-15
17. Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.-P., Meduri, E., Barrett, A., Nisbett, J., Sekowska, M., Wilk, A., Shin, S.-Y., Glass, D., Travers, M., Min, J.L., Ring, S., Ho, K., Thorleifsson, G., Kong, A., Thorsteindottir, U., Ainali, C., Dimas, A.S., Hassanali, N., Ingle, C., Knowles, D., Krestyaninova, M., Lowe, C.E., Di Meglio, P., Montgomery, S.B., Parts, L., Potter, S., Surdulescu, G., Tsaprouni, L., Tsoka, S., Bataille, V., Durbin, R., Nestle, F.O., O'Rahilly, S., Soranzo, N., Lindgren, C.M., Zondervan, K.T., Ahmadi, K.R., Schadt, E.E., Stefansson, K., Smith, G.D., McCarthy, M.I., Deloukas, P., Dermitzakis, E.T., Spector, T.D., Consortium, T.M.T.H.E.R.M.: Mapping cis- and trans-regulatory effects across multiple tissues in twins **44**(10), 1084–1089. doi:10.1038/ng.2394. Accessed 2015-10-15
18. Maksimovic, J., Gordon, L., Oshlack, A.: SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips **13**(6), 44. doi:10.1186/gb-2012-13-6-r44. Accessed 2015-10-15
19. Bolstad, B.M., Irizarry, R.A., Åstrand, M., Speed, T.P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias **19**(2), 185–193. doi:10.1093/bioinformatics/19.2.185. Accessed 2015-10-15

20. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing **57**(1), 289–300. Accessed 2015-10-15
21. Dedeurwaerder, S., Defrance, M., Calonne, E., Denis, H., Sotiriou, C., Fuks, F.: Evaluation of the infinium methylation 450k technology **3**(6), 771–784. doi:10.2217/epi.11.105. Accessed 2015-10-15
22. Breitling, L., Yang, R., Korn, B., Burwinkel, B., Brenner, H.: Tobacco-smoking-related differential DNA methylation: 27k discovery and replication **88**(4), 450–457. doi:10.1016/j.ajhg.2011.03.003. Accessed 2015-10-15
23. Chen, Y.-a., Lemire, M., Choufani, S., Butcher, D.T., Grafodatskaya, D., Zanke, B.W., Gallinger, S., Hudson, T.J., Weksberg, R.: Discovery of cross-reactive probes and polymorphic CpGs in the illumina infinium HumanMethylation450 microarray **8**(2), 203–209. doi:10.4161/epi.23470. Accessed 2015-10-15
24. Relton, C.L., Groom, A., St. Pourcain, B., Sayers, A.E., Swan, D.C., Embleton, N.D., Pearce, M.S., Ring, S.M., Northstone, K., Tobias, J.H., Trakalo, J., Ness, A.R., Shaheen, S.O., Davey Smith, G.: DNA methylation patterns in cord blood DNA and body size in childhood **7**(3), 31821. doi:10.1371/journal.pone.0031821. Accessed 2015-10-15
25. Wang, X., Zhu, H., Snieder, H., Su, S., Munn, D., Harshfield, G., Maria, B.L., Dong, Y., Treiber, F., Gutin, B., Shi, H.: Obesity related methylation changes in DNA of peripheral blood leukocytes **8**(1), 87. doi:10.1186/1741-7015-8-87. Accessed 2015-10-15
26. Almén, M.S., Jacobsson, J.A., Moschonis, G., Benedict, C., Chrousos, G.P., Fredriksson, R., Schiöth, H.B.: Genome wide analysis reveals association of a FTO gene variant with epigenetic changes **99**(3), 132–137. doi:10.1016/j.ygeno.2011.12.007. Accessed 2015-10-15
27. Feinberg, A.P., Irizarry, R.A., Fradin, D., Aryee, M.J., Murakami, P., Aspelund, T., Eiriksdottir, G., Harris, T.B., Launer, L., Gudnason, V., Fallin, M.D.: Personalized epigenomic signatures that are stable over time and covary with body mass index **2**(49), 49–674967. doi:10.1126/scitranslmed.3001262. Accessed 2015-10-15
28. Moleres, A., Campión, J., Milagro, F.I., Marcos, A., Campoy, C., Garagorri, J.M., Gómez-Martínez, S., Martínez, J.A., Azcona-Sanjulián, M.C., Martí, A.: Differential DNA methylation patterns between high and low responders to a weight loss intervention in overweight or obese adolescents: the EVASYON study **27**(6), 2504–2512. doi:10.1096/fj.12-215566. Accessed 2015-10-15
29. Rönn, T., Volkov, P., Davegårdh, C., Dayeh, T., Hall, E., Olsson, A.H., Nilsson, E., Tornberg, A., Dekker Nitert, M., Eriksson, K.-F., Jones, H.A., Groop, L., Ling, C.: A six months exercise intervention influences the genome-wide DNA methylation pattern in human adipose tissue **9**(6), 1003572. doi:10.1371/journal.pgen.1003572. Accessed 2015-10-15
30. Greer, S.N., Metcalf, J.L., Wang, Y., Ohh, M.: The updated biology of hypoxia-inducible factor: The updated biology of HIF **31**(11), 2448–2460. doi:10.1038/emboj.2012.125. Accessed 2015-10-15
31. Heikkilä, M., Pasanen, A., Kivirikko, K.I., Myllyharju, J.: Roles of the human hypoxia-inducible factor (HIF)-3alpha variants in the hypoxia response **68**(23), 3885–3901. doi:10.1007/s00018-011-0679-5. Accessed 2015-10-15
32. Makino, Y., Cao, R., Svensson, K., Bertilsson, G., Asman, M., Tanaka, H., Cao, Y., Berkenstam, A., Poellinger, L.: Inhibitory PAS domain protein is a negative regulator of hypoxia-inducible gene expression **414**(6863), 550–554. doi:10.1038/35107085. Accessed 2015-10-15
33. Makino, Y., Kanopka, A., Wilson, W.J., Tanaka, H., Poellinger, L.: Inhibitory PAS domain protein (IPAS) is a hypoxia-inducible splicing variant of the hypoxia-inducible factor-3alpha locus **277**(36), 32405–32408. doi:10.1074/jbc.C200328200. Accessed 2015-10-15
34. Jiang, C., Qu, A., Matsubara, T., Chanturiya, T., Jou, W., Gavrilova, O., Shah, Y.M., Gonzalez, F.J.: Disruption of hypoxia-inducible factor 1 in adipocytes improves insulin sensitivity and decreases adiposity in high-fat diet–fed mice **60**(10), 2484–2495. doi:10.2337/db11-0174. Accessed 2015-10-15
35. Shin, M.-K., Drager, L.F., Yao, Q., Bevans-Fonti, S., Yoo, D.-Y., Jun, J.C., Aja, S., Bhanot, S., Polotsky, V.Y.: Metabolic consequences of high-fat diet are attenuated by suppression of HIF-1alpha **7**(10), 46562. doi:10.1371/journal.pone.0046562. Accessed 2015-10-15
36. Zhang, H., Zhang, G., Gonzalez, F.J., Park, S.-m., Cai, D.: Hypoxia-inducible factor directs POMC gene to mediate hypothalamic glucose sensing and energy balance regulation **9**(7), 1001112. doi:10.1371/journal.pbio.1001112. Accessed 2015-10-15
37. Park, Y.S., David, A.E., Huang, Y., Park, J.-B., He, H., Byun, Y., Yang, V.C.: In vivo delivery of cell-permeable antisense hypoxia-inducible factor 1alpha oligonucleotide to adipose tissue reduces adiposity in obese mice **161**(1), 1–9. doi:10.1016/j.jconrel.2012.04.026. Accessed 2015-10-15
38. Heidbreder, M., Qadri, F., Jöhren, O., Dendorfer, A., Depping, R., Fröhlich, F., Wagner, K.F., Dominiak, P.: Non-hypoxic induction of HIF-3alpha by 2-deoxy-d-glucose and insulin **352**(2), 437–443. doi:10.1016/j.bbrc.2006.11.027. Accessed 2015-10-15
39. Hatanaka, M., Shimba, S., Sakaue, M., Kondo, Y., Kagechika, H., Kokame, K., Miyata, T., Hara, S.: Hypoxia-inducible factor-3alpha functions as an accelerator of 3t3-l1 adipose differentiation **32**(7), 1166–1172. doi:10.1248/bpb.32.1166
40. Robciuc, M.R., Naukkarinen, J., Ortega-Alonso, A., Tyynismaa, H., Raivio, T., Rissanen, A., Kaprio, J., Ehnholm, C., Jauhiainen, M., Pietiläinen, K.H.: Serum angiopoietin-like 4 protein levels and expression in adipose tissue are inversely correlated with obesity in monozygotic twins **52**(8), 1575–1582. doi:10.1194/jlr.P015867. Accessed 2015-10-15
41. Bell, J.T., Pai, A.A., Pickrell, J.K., Gaffney, D.J., Pique-Regi, R., Degner, J.F., Gilad, Y., Pritchard, J.K.: DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines **12**(1), 10. doi:10.1186/gb-2011-12-1-r10. Accessed 2015-10-15
42. Gibbs, J.R., van der Brug, M.P., Hernandez, D.G., Traynor, B.J., Nalls, M.A., Lai, S.-L., Arepalli, S., Dillman, A., Rafferty, I.P., Troncoso, J., Johnson, R., Zielke, H.R., Ferrucci, L., Longo, D.L., Cookson, M.R., Singleton, A.B.: Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain **6**(5), 1000952. doi:10.1371/journal.pgen.1000952. Accessed 2015-10-15
43. Bonsignore, M.R., McNicholas, W.T., Montserrat, J.M., Eckel, J.: Adipose tissue in obesity and obstructive

sleep apnoea **39**(3), 746–767. doi:10.1183/09031936.00047010. Accessed 2015-10-15
44. Cockerill, P.N.: Structure and function of active chromatin and DNase i hypersensitive sites **278**(13), 2182–2210. doi:10.1111/j.1742-4658.2011.08128.x. Accessed 2015-10-15
45. Kraft, P.: Curses—winner's and otherwise—in genetic epidemiology: **19**(5), 649–651. doi:10.1097/EDE.0b013e318181b865. Accessed 2015-10-15

2 Association entre méthylation de l'ADN au locus *CPT1A* et les taux plasmatiques de lipides

Les taux plasmatiques de lipides sont des facteurs de risque héritables (et accessoirement modifiables) des pathologies cardiovasculaires. Entre 100 à 150 polymorphismes sont connus pour influencer les taux lipidiques et expliquer de 10 à 12% de la variabilité des taux (Global Lipids Genetics Consortium 2013 ; Teslovich *et al.* 2010).

Dans le but de mieux expliquer la variabilité des taux lipidiques plasmatiques, j'ai entrepris une analyse du méthylome sanguin via une approche MWAS dans l'étude MARTHA. Au même moment, une autre équipe publiait les résultats d'une étude similaire à celle que j'étais en train d'effectuer (Frazier-Wood *et al.* 2014) où était trouvé que la méthylation de deux sites CpG (cg00574958 et cg17058475) du locus *CPT1A* était associée aux taux de lipoprotéines de basse densité (LDL) et de lipoprotéines de très basse densité (VLDL). Je me suis donc intéressé à ces deux sites CpG et j'ai pu confirmer qu'ils étaient également associés au taux de triglycérides dans l'étude MARTHA (Gagnon *et al.* 2014) avec une p-value de $8,28 * 10^{-6}$ pour le site cg00574958 et de $8,86 * 10^{-3}$ pour le site cg17058475. J'ai ensuite retrouvé ces associations dans l'étude F5L-Pedigrees avec une p-value de $3,28 * 10^{-3}$ pour le site cg00574958 et de $2,88 * 10^{-3}$ pour le site cg17058475. Une augmentation de 0,1 du niveau de méthylation du site CpG cg00574958 est associée à une diminution de 0,059 du taux de triglycéride (log-transformé) dans l'étude MARTHA et d'une diminution de 0,054 dans l'étude F5L-Pedigrees. Tandis qu'une augmentation de 0.1 du niveau de méthylation du site cg00574958 est associée à une diminution de 0,025 du taux de triglycéride (log-transformé) dans la première étude et d'une diminution de 0,041 dans la seconde étude. Ces résultats ont permis de confirmer, malgré l'absence de véritable réplication dans le travail de Frazier-Wood *et al.* 2014, que les niveaux de méthylation du locus *CPT1A* sont bien associés aux taux de lipides.

Ces travaux montrent en outre qu'il est possible de détecter à partir d'échantillons sanguins des marques de méthylation de l'ADN d'une enzyme exprimée dans le foie et les glandes tissulaires. La validation est d'autant plus robuste que les individus des différentes études sont d'origines différentes (d'origine marseillaise pour MARTHA, franco-canadienne pour F5L-Pedigrees et de Minneapolis et Salt Lake City pour l'étude GOLDN de Frazier-Wood *et al.* 2014). De plus, la quantification des niveaux de méthylation de l'ADN a été réalisée par des méthodes de normalisation qui différaient entre les données de l'étude GOLDN et celles que j'avais à ma disposition (voir article ci-après) et

les lipides ont été mesurés avec des techniques différentes (spectrophotométrie pour MARTHA et F5L-Pedigrees et spectroscopie de résonance magnétique nucléaire pour GOLDN).

Cette fois encore, l'association entre le locus *CPT1A* et le taux plasmatique de lipides n'a pas été détectée par l'approche GWAS sur environ 190 000 sujets (Global Lipids Genetics Consortium 2013) alors qu'elle est retrouvée dans nos travaux MWAS portant sur seulement 526 individus et confirmée par la suite dans d'autres études MWAS réalisées seulement sur 991 (Irvin *et al.* 2014), 48 (Tobi *et al.* 2014) et 6113 sujets (Demerath *et al.* 2015). Ces résultats démontrent une nouvelle fois l'intérêt de l'approche MWAS comme une des stratégies pertinentes pour identifier de nouveaux mécanismes associés à la variabilité inter-individuelle de biomarqueurs du risque cardiovasculaire. L'ensemble de ces travaux est décrit dans la publication ci-après.

En complément de ce travail, j'ai réalisé des méta-analyses à effets aléatoires des MWAS réalisées sur MARTHA et F5L-Pedigrees indépendamment sur les taux de lipoprotéines de basse densité (LDL), lipoprotéines de haute densité (HDL) et triglycérides (TG). La méta-analyse sur les taux de HDL a mis en évidence une association significative ($p = 3,31 * 10^{-9}$) pour le site CpG cg06500161 du gène *ABCG1* (figure 5.1). Cette association entre la méthylation de l'ADN du gène *ABCG1* et le taux plasmatique de HDL est également trouvée dans une autre étude de 1776 sujets (Pfeiffer *et al.* 2015).

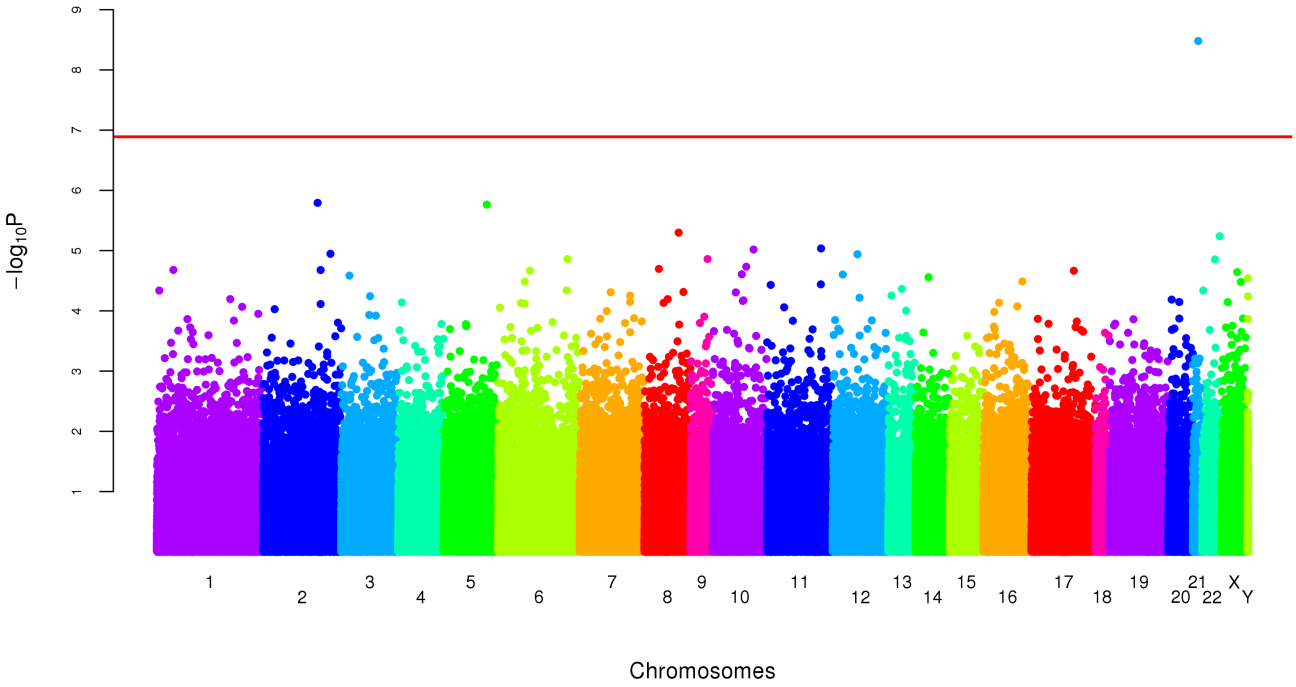

FIGURE 5.1 – Graphique de type Manhattan des résultats d'une méta-analyse à effets aléatoires des études MARTHA et F5L-Pedigrees sur le HDL. Le point bleu situé au-dessus de la ligne rouge (seuil de Bonferroni) correspond au site CpG cg06500161 du gène *ABCG1* sur le chromosome 21.

Aucun site CpG ne passe le seuil de Bonferroni pour les LDL. Pour les TG, en plus du locus *CPT1A*, nous retrouvons le même site CpG de *ABCG1* avec une p-value de $4,81 * 10^{-5}$ et un second site CpG cg27243685 du même gène avec une p-value plus forte $1,24 * 10^{-7}$. Ces résultats n'ont pas été publiés car une autre équipe nous a précédé (PFEIFFER *et al.* 2015).

COMMENTARY

Robust validation of methylation levels association at CPT1A locus with lipid plasma levels

France Gagnon[1], Dylan Aïssi[2,3,4], Alain Carrié[3,4,5], Pierre-Emmanuel Morange[6,7,8] and David-Alexandre Trégouët[2,3,4*]

*Correspondence:
david.tregouet@upmc.fr
[2] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, F-75013 Paris, France
[3] INSERM, UMR_S 1166, F-75013 Paris, France
[4] ICAN Institute for Cardiometabolism and Nutrition, F-75013 Paris, France
Full list of author information is available at the end of the article

NOTE:
This file is an Author's Post-print version using the BioMed Central TeX template.
For the Publisher's Version/PDF, please see :
The Journal of Lipid Research, 2014 May 21;55(7):1189-1191.
DOI: 10.1194/jlr.E051276.

There is increasing enthusiasm regarding the use of biobanked whole blood DNA as a model to discover methylation marks associated with biological phenotypes and generate novel mechanistic hypotheses [1, 2, 3]. DNA methylation has a critical role in cell functions and is cell-type specific. Such cell specificity makes DNA methylation particularly challenging for epidemiological epigenetic investigations because disease relevant cell types might not be accessible due to practical issues such as availability, ethics, and cost associated with more complex specimen collection.

Recent work suggests that agnostic methylation-wide association scan (MWAS) in peripheral blood can reflect phenotype-associated methylation marks in other tissues and cell types, with effects detected in established effector cells much stronger than effects detected in blood [4]. These observations suggest that marks detected in blood are associated with functions in effector cells. The Illumina HumanMethylation450 (HM450K) array is a robust assay to measure DNA methylation across the genome [4, 5, 6, 7]. For any high-throughput technologies, and in particular for a novel assay such as the HM450K, rigorous quality control procedures are warranted and robustness of findings must be validated through independent replication to avoid reporting spurious associations.

In the current issue of the Journal, Frazier-Wood et al. [8] reported the novel findings of significant negative correlations between methylation levels at two CpG sites in the *CPT1A* locus and plasma levels of VLDL and LDL. Methylation levels were assessed in CD4+ T-cells isolated from peripheral blood DNA using the HM450K array. Given that no independent study samples were available for replication, to circumvent this challenge, the authors adopted an internal validation method by splitting the whole sample into "discovery and replication subsamples". This strategy provides arguments in favor of the discovered associations but does not provide evidence of robustness against spurious findings due to sampling or confounding

biases or any other undetected biases present in the study sample. A robust and thorough validation strategy implies the use of independent study samples and variation in the study designs [9, 10]. The validation phase is of particular importance in MWAS, as this technique is particularly subjected to confounders [3]. Thus, we undertook to test for associations the two CTP1A CpG sites found associated with lipid-related traits by Frazier-Wood et al. using two independent study samples with considerable variations in their respective study design and with the design of the Frazier-Wood study.

The studies had differences in sampling scheme, DNA methylation specimen, and array preprocessing approaches. The notable differences in the design and sample characteristics between the three studies are shown in Table 1. Most notable is the method for lipid measurement, nuclear magnetic resonance spectroscopy in Frazier-Wood et al. and spectrophotometry in our studies. In addition to sampling variation, and of particular interest for MWAS studies, Frazier-Wood et al. assessed DNA methylation in isolated CD4+ T-cells, while we assessed methylation in peripheral whole blood, which includes CD4+ T-cell ($< 30\%$) and several other leukocyte subtypes. Finally, different normalization procedures were used: we applied the SWAN methodology [4, 11] to globally normalize β values from the Infinium I and II probes, while separate normalization by probe type was applied by Frazier-Wood et al.

Despite the nontrivial differences between these studies, we observed strong statistical evidence for a negative association between the two CTP1A CpG sites (cg00574958 and cg17058475) identified by Frazier-Wood et al. and plasma levels of both LDL and triglycerides (TG) in two independent studies, the MARTHA [4, 12] and F5L-pedigree studies [13]. In our two samples totaling 526 individuals, increased DNA methylation levels at CPTA1 CpG sites were associated with both decreased LDL and TG (Table 2). A 1% increase in cg00574958 DNA methylation levels was associated with a 0.057 ± 0.011 decrease in log TG levels ($P = 5.71 * 10^{-8}$). Corresponding values for a 1% increase in cg17058475 levels were 0.030 ± 0.008 ($P = 9.83 * 10^{-5}$).

Of note, cg00574958 and cg17058475 were highly correlated ($\rho_{spearman} = 0.67$ in both studies, $P < 10^{-16}$); adjusting for cg00574958 in the model abolished the effect observed for cg17058475 on log TG levels. Finally, after adjustment for key covariates (age, sex, BMI, cell type composition, batch, and chip effects), cg00574958 explained $\sim 4\%$ of log TG plasma levels, both in MARTHA and F5L-pedigrees. Negative association was also observed between plasma LDL levels and cg17058475 ($P = 1.710^{-2}$) but not with cg00574958 ($P = 0.11$). No association was observed with HDL-cholesterol levels ($P = 0.96$ for cg00574958 and $P = 0.75$ for cg17058475), nor with total cholesterol levels ($P = 0.16$ for cg00574958 and $P = 0.53$ for cg17058475).

Table 1 Main design and sample characteristics of the three MWAS studies on lipids

Study name	MARTHA	F5L-Pedigrees	GOLDN (Frazier-Wood et al.)	
Study design	Unrelated individuals	Extended pedigrees	Extended pedigrees	
Subjects Origin	Caucasians from Marseille area (South of France)	French-Canadians from Ottawa area (Canada)	European descent from Minneapolis (Minnesota) and Salt Lake City (Utah)	
			Discovery	Validation
N	327	199	663	331
Age	44.1 (14.23)	39.6 (16.9)	48.6 (16.4)	47.7 (16.6)
Sex (% male)	21.7	46.7	47	49.2
Total cholesterol	5.452 (1.019) (g/L)	4.896 (1.079) (g/L)	NA	NA
HDL-cholesterol	1.476 (0.435) (g/L)	1.359 (0.353) (g/L)	40.0 (5.6) (nmol/L)	37.0 (5.8) (nmol/L)
LDL-cholesterol	3.647 (0.980) (g/L)	3.111 (0.901) (g/L)	1393.8 (460.0) (nmol/L)	1369.5 (1369.5) (nmol/L)
Triglycerides	1.058 (0.772) (mmol/L)	1.487 (0.905) (mmol/L)	NA	NA
Lipid measurement technology	Spectrophotometry except for LDL that was derived from the Friedewald's formula		Nuclear Magnetic Resonance spectroscopy	
Blood collection	Fasting	Fasting and non smoking	Fasting	
DNA specimen	Whole blood		Isolated CD4+ T-cells	
Medication	No exclusion	Exclusion if on medication	Asked to discontinue the use of lipid lowering drugs and over-the-counter medication that could affect lipid levels.	
HumanMethylation450k Normalization	Noob[a] and SWAN[b]		Separately normalized probes from the Infinium I and II using ComBat[c]	
Adjustment	Age, sex, batch effect, chip effect, cell type composition, dyslipidemia	Age, sex, batch effect, chip effect, cell type composition[d], family structure	Age, sex, study site, T-cell purity (based on the first 4 principal components), family structure	

[a] Triche et al. 2013. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 41(7):e90. [14]

[b] Maksimovic et al. 2012. SWAN: Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. Genome Biol. 13(6):R44. [11]

[c] Johnson et al. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 8(1):118-27. [15]

[d] In MARTHA, specific measured biological counts of lymphocytes, monocytes, neutrophils, eosinophils and basophils were used to characterize leukocytes composition. In F5L-pedigrees, adjustment for cell type composition was handled by the methods described in Houseman et al. [16].

Table 2 Association of cg00574958 and cg17058475 CPT1A CpG variability with plasma TG and LDL levels in the MARTHA and F5L-pedigrees

		TG (log)	LDL
cg00574958	MARTHA	-0.059 (0.013) $P = 8.28 * 10^{-6}$	-0.023 (0.040) $P = 0.57$
	F5L-Pedigrees	-0.054 (0.018) $P = 3.28 * 10^{-3}$	-0.046 (0.019) $P = 0.12$
	Combined [a]	-0.057 (0.011) $P = 5.71 * 10^{-8}$	-0.038 (0.024) $P = 0.11$
cg17058475	MARTHA	-0.025 (0.009) $P = 8.86 * 10^{-3}$	-0.051 (0.029) $P = 8.57 * 10^{-2}$
	F5L-Pedigrees	-0.041 (0.013) $P = 2.88 * 10^{-3}$	-0.037 (0.022) $P = 9.36 * 10^{-2}$
	Combined [a]	-0.030 (0.008) $P = 9.83 * 10^{-5}$	-0.042 (0.017) $P = 1.7 * 10^{-2}$

Association was tested using a linear regression model (mixed linear model in F5L-Pedigrees) where log(TG) (LDL, resp.) was the outcome and the CpG site the predictor variable. Analyses were adjusted for age, sex, cell type, batch and chip effects. Reported coefficients (standard error) represent the increase in outcome value associated with a 1% increase in CpG site variability. In MARTHA, TG and LDL phenotypes were measured in 327 and 180 individuals, respectively. In the F5L-pedigrees study, lipid phenotypes were measured in 199 individuals.
[a] Results of the MARTHA and F5L-pedigrees studies were combined into a random effect meta-analysis based on the inverse-variance weighting method.

The CPT1A protein is essential for fatty acid oxidation (a multistep process that metabolizes fats and converts them into energy) and is expressed in the liver and glandular tissues [17]. This pivotal role in fatty acid metabolism makes *CPT1A* DNA methylation marks relevant to many metabolic disorders (from lipids to glucose homeostasis). The lipid-related DNA methylation probes in this study (cg00574958 and cg17058475) are designated as falling in a single "CpG shore", and are flanked by two CpG islands. Human ENCODE HM450K studies performed on over 40 cell lines suggest these two probes show more variable methylation levels than the two CpG islands that flank them. The uncoupled methylation levels at these probes versus the flanking islands suggest that the observed variation is more likely to be regulatory. This region also shows evidence of open chromatin through DNase I hypersensitivity assays [18] and gene regulatory potential through chromatin immunoprecipitation sequencing of the epigenetic modification H3K27ac [19]. More work is needed to understand the functional impact of DNA methylation on *CPT1A* gene regulation.

Three important conclusions emerge from this validation study. First, despite limitations in the Frazier-Wood et al. replication approach, the published results are robust to variation in sample, study design, normalization procedures, and even DNA blood specimen type. Second, inter-individual variation in lipid-related traits appears to be under the influence of DNA methylation regulation at the *CPT1A* locus. This epidemiological evidence now requires technical validation and functional work to confirm that these methylation marks are causes rather than consequences of lipid levels variation. Given that DNA methylation marks are potentially reversible, evidence for their role in the regulation of such a key enzyme is of great interest as it could lead to new therapeutic approaches (e.g., drug and/or diet supplementation) to modulate *CPT1A* expression. Finally, and of major importance for MWAS studies, peripheral whole blood DNA methylation marks were detected in an enzyme gene expressed in the liver and glandular tissues, suggesting that such marks could serve as surrogates for methylation at more closely-related effector cells, such as hepatocytes. The latter adds to the recent paper by Dick et al. [4] also supporting the value of peripheral whole blood DNA methylation marks as biomarkers of methylation in other tissues.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
We thank Dr. Michael D. Wilson for his judicious comments on the manuscript and for the many fruitful discussions about epigenetic regulation.
D.A. was supported by a PhD grant from the Région Ile de France (CORDDIM). The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique. The F5L Thrombophilia French-Canadian Pedigree study was supported by grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). The Human450Meethylation epityping was partially funded by the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (grant T6484). F.G. holds a Canada Research Chair. Statistical analyses were performed using the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05).

Author details
[1]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [2] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, F-75013 Paris, France. [3] INSERM, UMR_S 1166, F-75013 Paris, France. [4] ICAN Institute for Cardiometabolism and Nutrition, F-75013 Paris, France. [5] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team Integrative Biology of Atherosclerosis, F-75013 Paris, France. [6] Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France. [7] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France. [8] Laboratory of Haematology, La Timone Hospital, F-13385 Marseille, France.

References

1. Murphy, T.M., Mill, J.: Epigenetics in health and disease: heralding the EWAS era **383**(9933), 1952–1954. doi:10.1016/S0140-6736(14)60269-5. Accessed 2015-10-14
2. Osório, J.: Obesity: Looking at the epigenetic link between obesity and its consequences—the promise of EWAS **10**(5), 249–249. doi:10.1038/nrendo.2014.42. Accessed 2015-10-14
3. Callaway, E.: Epigenomics starts to make its mark **508**(7494), 22–22. doi:10.1038/508022a. Accessed 2015-10-14
4. Dick, K.J., Nelson, C.P., Tsaprouni, L., Sandling, J.K., Aïssi, D., Wahl, S., Meduri, E., Morange, P.-E., Gagnon, F., Grallert, H., Waldenberger, M., Peters, A., Erdmann, J., Hengstenberg, C., Cambien, F., Goodall, A.H., Ouwehand, W.H., Schunkert, H., Thompson, J.R., Spector, T.D., Gieger, C., Trégouët, D.-A., Deloukas, P., Samani, N.J.: DNA methylation and body-mass index: a genome-wide analysis **383**(9933), 1990–1998. doi:10.1016/S0140-6736(13)62674-4. Accessed 2015-03-26
5. Zeilinger, S., Kühnel, B., Klopp, N., Baurecht, H., Kleinschmidt, A., Gieger, C., Weidinger, S., Lattka, E., Adamski, J., Peters, A., Strauch, K., Waldenberger, M., Illig, T.: Tobacco smoking leads to extensive genome-wide changes in DNA methylation **8**(5), 63812. doi:10.1371/journal.pone.0063812. Accessed 2015-04-22
6. Bell, J.T., Tsai, P.-C., Yang, T.-P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., Shin, S.-Y., Dempster, E.L., Murray, R.M., Grundberg, E., Hedman, A.K., Nica, A., Small, K.S., Dermitzakis, E.T., McCarthy, M.I., Mill, J., Spector, T.D., Deloukas, P., The MuTHER Consortium: Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population **8**(4), 1002629. doi:10.1371/journal.pgen.1002629. Accessed 2015-10-14
7. Dayeh, T., Volkov, P., Salö, S., Hall, E., Nilsson, E., Olsson, A.H., Kirkpatrick, C.L., Wollheim, C.B., Eliasson, L., Rönn, T., Bacos, K., Ling, C.: Genome-wide DNA methylation analysis of human pancreatic islets from type 2 diabetic and non-diabetic donors identifies candidate genes that influence insulin secretion **10**(3), 1004160. doi:10.1371/journal.pgen.1004160. Accessed 2015-10-14
8. Frazier-Wood, A.C., Aslibekyan, S., Absher, D.M., Hopkins, P.N., Sha, J., Tsai, M.Y., Tiwari, H.K., Waite, L.L., Zhi, D., Arnett, D.K.: Methylation at CPT1a locus is associated with lipoprotein subfraction profiles **55**(7), 1324–1330. doi:10.1194/jlr.M048504. Accessed 2015-04-30
9. Rosenbaum, P.R.: Replicating effects and biases **55**(3), 223–227. doi:10.1198/000313001317098220. Accessed 2015-10-14
10. Kraft, P., Zeggini, E., Ioannidis, J.P.A.: Replication in genome-wide association studies **24**(4), 561–573. doi:10.1214/09-STS290. Accessed 2015-10-14
11. Maksimovic, J., Gordon, L., Oshlack, A.: SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips **13**(6), 44. doi:10.1186/gb-2012-13-6-r44. Accessed 2015-10-14
12. Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Kallel, C., Zelenika, D., Lathrop, M., Trégouët, D.-A., Morange, P.-E.: Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein c anticoagulant pathway: the MARTHA project **157**(2), 230–239. doi:10.1111/j.1365-2141.2011.09025.x. Accessed 2015-10-14
13. Antoni, G., Morange, P.-E., Luo, Y., Saut, N., Burgos, G., Heath, S., Germain, M., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Zelenika, D., Alessi, M.-C., Drouet, L., Visvikis-Siest, S., Wells, P.S., Lathrop, M., Emmerich, J., Tregouet, D.-A., Gagnon, F.: A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism **8**(12), 2671–2679. doi:10.1111/j.1538-7836.2010.04092.x. Accessed 2015-10-14
14. Triche, T.J., Weisenberger, D.J., Berg, D.V.D., Laird, P.W., Siegmund, K.D.: Low-level processing of illumina infinium DNA methylation BeadArrays **41**(7), 90–90. doi:10.1093/nar/gkt090. Accessed 2015-05-29
15. Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods **8**(1), 118–127. doi:10.1093/biostatistics/kxj037. Accessed 2015-03-26
16. Houseman, E.A., Accomando, W.P., Koestler, D.C., Christensen, B.C., Marsit, C.J., Nelson, H.H., Wiencke, J.K., Kelsey, K.T.: DNA methylation arrays as surrogate measures of cell mixture distribution **13**(1), 86. doi:10.1186/1471-2105-13-86. Accessed 2015-03-26

17. Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., Ponten, F.: Towards a knowledge-based human protein atlas **28**(12), 1248–1250. doi:10.1038/nbt1210-1248. Accessed 2015-10-14
18. Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A.: The accessible chromatin landscape of the human genome **489**(7414), 75–82. doi:10.1038/nature11232. Accessed 2015-10-14
19. Consortium, T.E.P.: An integrated encyclopedia of DNA elements in the human genome **489**(7414), 57–74. doi:10.1038/nature11247. Accessed 2015-10-14

3 Recherche de profils de méthylation associés à la génération de thrombine

Au cours de ma thèse, j'ai également participé à un travail de thèse d'une autre doctorante de l'équipe : Ares Rocañín Arjó. L'objectif du projet d'Ares Rocañín Arjó était d'identifier des déterminismes génétiques et épigénétiques de phénotypes quantitatifs associés à la génération de thrombine (Rocañín-Arjó *et al.* 2015).

La prothrombine (facteur de coagulation II ou FII) est synthétisée dans le foie sous une forme zymogène. Une fois relâchée puis clivée, cette dernière va exposer ses domaines et activer ses fonctions pour devenir la thrombine (facteur de coagulation II activé ou FIIa). La thrombine a un rôle central dans la cascade de coagulation. Elle sert de catalyseur à la formation de la fibrine (protéine qui formera le caillot sanguin), d'activateur pour d'autres facteurs pro-coagulants ainsi que de déclencheur pour des d'enzymes anti-coagulantes. Son activation résulte de la cascade de coagulation à laquelle participent de nombreuses protéines qui s'activent les unes les autres (figure 5.2).

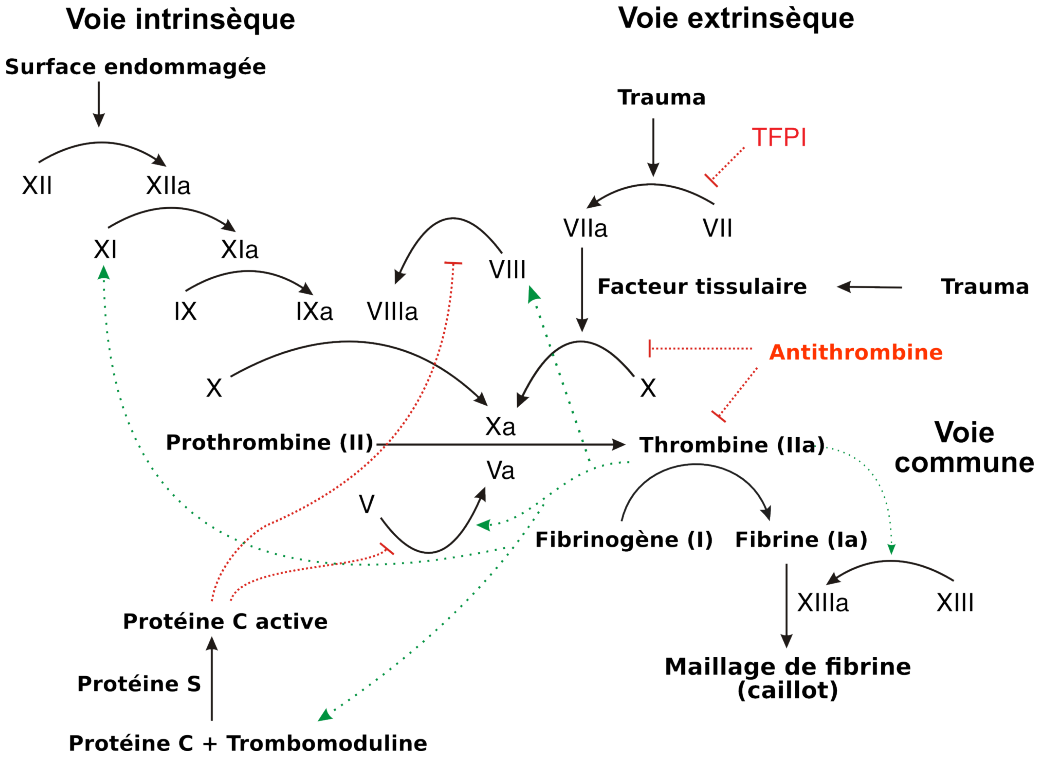

Figure 5.2 – Cascade de coagulation et les différentes protéines impliquées.

La thrombine participe également à des systèmes physiologiques autres que l'hémostase (les systèmes immunitaire, nerveux, gastro-intestinal et musculo-squelettique). Elle interagit également avec

des protéines et des récepteurs, en jouant sur l'activation des plaquettes et des cellules endothéliales ainsi que sur la stimulation de l'adhésion, l'angiogenèse, la croissance cellulaire, la différenciation, la prolifération, la vasoconstriction et l'inflammation. La combinaison complexe de substrats et de cofacteurs, permet de moduler ce large éventail de fonctions. Une régulation fine de la thrombine est fondamentale pour contrôler toutes ces fonctions et ainsi assurer une physiologie normale. En cas de déséquilibre des niveaux de thrombine, différents types d'anomalies peuvent se produire dont les plus connues et étudiées sont la thrombose, l'hémophilie, l'inflammation et l'athérosclérose.

La mesure très précise de son activité est alors indispensable pour étudier et comprendre les conséquences physiopathologiques et moléculaires de ses anomalies. Le test du potentiel de génération de thrombine (TGP) est un test dont le but est de mesurer la quantité potentielle de thrombine qui est capable d'être activée au moment de coagulation. Ses résultats reflètent l'ensemble du processus de la coagulation à partir de l'initiation, la propagation et la terminaison-amplification.

Il est possible d'extraire plusieurs paramètres quantitatifs de ce test, dont les plus utilisés sont : le temps de latence (*lagtime*) mesurant le temps écoulé depuis l'initialisation de la coagulation jusqu'au début de la formation du caillot ; le potentiel endogène de la thrombine (*ETP* pour *Endogeneous Thrombin Potential*) représentant la quantité totale de la thrombine activée ce qui permet de représenter au mieux l'état de coagulation et tout le travail enzymatique autour de la génération de la thrombine ; la hauteur du pic (*peak*) mesurant la quantité maximale de thrombine activée au cours du processus.

J'ai assisté Ares Rocañín Arjó dans son étude visant à identifier, à partir de l'ADN sanguin, des marques de méthylation de l'ADN associées à ces trois biomarqueurs plasmatiques de la génération de thrombine. À partir des niveaux de méthylation de l'ADN du sang périphérique des études MARTHA et F5L-Pedigrees, une approche MWAS a été réalisée pour chacun des trois biomarqueurs. Les détails de cette analyse sont décrits dans la publication (Rocañín-Arjó *et al.* 2015) jointe à la fin de cette section. Aucun des sites CpG n'a été retrouvé significativement associé de manière consistante dans les deux échantillons aux phénotypes étudiés. L'absence d'associations peut s'expliquer par plusieurs hypothèses. Soit l'ADN provenant du sang périphérique n'est pas un bon modèle pour identifier des sites CpG dont le niveau de méthylation est associé à la génération de thrombine. Soit les associations entre marques de méthylation observables dans le sang et génération de thrombine sont trop faibles pour être identifiées avec les échantillons disponibles dans cette étude. Il serait alors nécessaire d'avoir recours à de plus grandes cohortes pour avoir plus de puissance statistique et ainsi les détecter (Tsai

& J. T. Bell 2015).

LETTER TO THE EDITORS-IN-CHIEF

Thrombin Generation Potential and Whole-Blood DNA methylation

Ares Rocanin-Arjo[1,2,3], Jessica Dennis[4], Pierre Suchon[5,6,7], Dylan Aïssi[1,2,3], Vinh Truong[4], David-Alexandre Trégouët[1,2,3], France Gagnon[4] and Pierre-Emmanuel Morange[5,6,7*]

*Correspondence: pierre.morange@ap-hm.fr
[5] Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France
[6] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France
[7] Laboratory of Haematology, La Timone Hospital, F-13385 Marseille, France
Full list of author information is available at the end of the article

NOTE:
This file is an Author's Post-print version using the BioMed Central TeX template.
For the Publisher's Version/PDF, please see :
Thrombosis Research, 2015 Mar;135(3):561-4.
DOI: 10.1016/j.thromres.2014.12.010.

Introduction

A complex cascade of coagulation proteins underlies hemostasis and prevents life-threatening blood loss from damaged blood vessels. Determining the inter-individual variability in plasma levels of the collective effect of these proteins is of great importance. In the last decade, agnostic genome-wide association studies have contributed to discovering novel genes and pathways associated with biomarkers of the coagulation cascade. However, the identified genetic factors account for a minor proportion of the heritability of these biomarkers, suggesting that other contributing factors remain to be identified. DNA methylation is one of the compelling sources that could contribute to the so-called missing heritability of quantitative biological traits. DNA methylation is an epigenetic mechanism that participates in the regulation of gene expression, generally through gene silencing. Several key molecules participating in the coagulation pathway, such as von Willebrand Factor [1], factor VII [2] and factor VIII [3], have been shown to be under the influence of DNA methylation marks. While the former finding was observed in endothelial and kidney cells, the latter two were derived from the analysis of DNA methylation measured in blood.

Intense discussions are emerging about the use of whole blood as a model tissue to discover methylation marks associated with biological phenotypes through methylation-wide association studies (MWAS) [4]. In this context, we explored whether the whole blood DNA based MWAS strategy may help discover novel mechanisms associated with hemostasis regulation. We applied this strategy to quantitative biomarkers for thrombotic disorders characterizing the overall hemostatic status of individuals as evaluated by a global thrombin generation assay [5].

Materials and Methods

MARTHA Study

We measured genome-wide DNA methylation in whole blood samples of 238 individuals of the MARTHA study [6] using the dedicated Illumina HumanMethylation450

array. A detailed description of the quality controls and the normalization procedures applied to the methylation array data has previously been published [7].

MARTHA subjects were assessed for a thrombin generation assay using the calibrated automated thrombography method [5] as extensively described in [8]. Three biological biomarkers were derived from the thrombin generation potential (TGP) measurements produced by the assay: the endogenous thrombin potential (ETP),the lag time and peak height (Supplementary Table 1).

A total of 388,120 CpG sites were then tested for association with three TGP biomarkers. For each TGP biomarker, association analyses were performed using a linear regression approach and adjusted for age, sex, oral contraception therapy, body-mass index, batch and chip effects [7] and cell type composition determined by specific biological counts of lymphocytes, monocytes, neutrophils, eosinophils and basophils.

Any association that was genome-wide significant at the Bonferroni threshold of $1.29 * 10^{-7}$ $(= 0.05/388,120)$ was sought for independent replication in the F5L pedigree study.

F5L Pedigree Study

A total of 187 related individuals of the F5L pedigree study described in [9] were phenotyped for the three TGP biomarkers (Supplementary Table 1) with exactly the same protocol and by the same laboratory; and simultaneously processed for the DNA methylation array with the MARTHA samples. Association analysis of CpG sites with TGP phenotypes was performed using a linear mixed effect model accounting for the familial relatedness and adjusted for the same factors as in the MARTHA study. As cell counts were not available in the F5L pedigree study, we used the method described in [10] to adjust for cell type composition.

Finally, a MWAS investigation was also carried out in the F5L pedigree study and the results were combined to those obtained in MARTHA through a random-effect meta-analysis using the GWAMA program [11].

Results and Conclusions

None of the tested 388,120 CpG sites was significantly associated with ETP or Lag Time at the Bonferroni threshold of $1.29 * 10^{-7}$. Conversely, one CpG, cg26285502, which mapped to *C15orf41* on chromosome 15q14, reached genome-wide significance for association with Peak height ($P = 7.15 * 10^{-8}$). A 1% increase in cg26285502 DNA methylation was associated with an 11.96 ± 2.13 nmol L^{-1} increase (increase $\pm$ SE) in adjusted Peak values. We sought to replicate this finding in an independent sample of 187 individuals of the F5L pedigree study. In F5L, a 1% increase in cg26285502 DNA methylation was associated with a non significant ($P = 0.40$) increase of 2.16 ± 2.57 nmol L^{-1} increase in adjusted Peak values.

In order to increase statistical power to detect variability in CpG sites associated with TGP biomarkers, we conducted a MWAS on each TGP biomarker in the F5L-pedigree study and combined the results with those obtained in MARTHA. A Manhattan plot representation of these results is shown in Fig. 1. No novel CpG association signal emerged at the pre-specified genome-wide threshold. The smallest observed p-values were $P = 2.44 * 10^{-6}$ (*SDAD1* cg27589582), $P = 5.8710^{-6}$

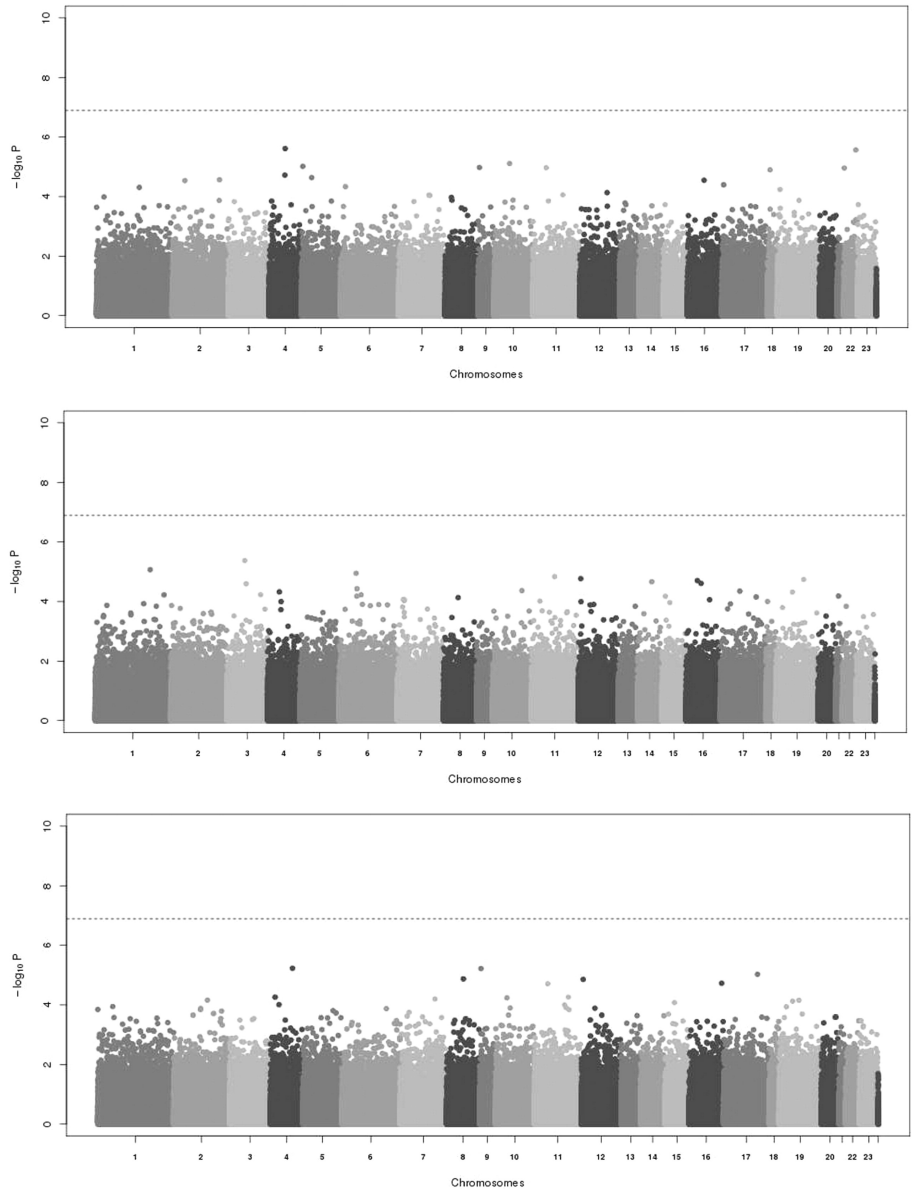

Figure 1 Meta MWAS results of MARTHA and F5L pedigrees.
Manhattan plot representation of the results derived from the meta-analysis of 2 methylome-wide association studies assessing the influence of DNA methylation levels at 388,120 CpG sites on TGP biomarkersin 425 individuals. ETP (Top), peak height (Middle), and lag time (Bottom). The horizontal line corresponds to the genome-wide significant threshold fixed at $1.29 * 10^{-7}$.

(cg18197092 mapping to a region with no gene on chromosome 4q26) and $P = 4.2310^{-6}$ (*MITF* cg06070625) for ETP, Peak and Lag Time, respectively.

Recently, it has been suggested that decreased levels of leukocyte DNA methylation at *NOS3* and *EDN1* genes could associate with higher ETP [12]. We did not observe strong statistical support of these findings in our combined datasets. Indeed, the *NOS3* CpG site showing the strongest association with ETP was the cg03469471 with combined $p - value = 0.165$. For the *EDN1* locus, the strongest

association was observed with cg07714708 ($p = 0.031$). However, this association was not consistent with the results observed in [12]. In our samples, a 1% increase in cg03469471 DNA methylation was associated with a 5.42 ± 2.50 increase in ETP (7.84 ± 3.77 and 3.51 ± 3.35 in MARTHA and F5L studies, respectively).

To our knowledge, this work is the first large-scale epidemiological investigation of DNA methylation marks measured in whole blood in relation to quantitative traits of the coagulation cascade. Even though DNA methylation measured in blood reflects average DNA levels from different cell types, the application of the MWAS strategy in whole blood cells has been proposed to detect differentially methylated regions (DMRs). Coagulation-relevant DNA methylation marks detected in blood could reflect stronger effects from effector cells [7], such as hepatocytes and endothelial cells. Such an approach may be particularly suited when access to relevant cell types/tissues is still not feasible at a large epidemiological scale.

We observed a genome-wide significant association of DNA methylation levels at *C15orf41* with Peak height variability in the MARTHA study but did not validate it in our replication study. *C15orf41* was an interesting candidate as it has very recently been found mutated in patients with congenital dyserythropoietic anaemia [13]. We investigated whether *C15orf41* methylation levels could associate with biological criteria of anemia (red blood cells counts and hemoglobin levels) in MARTHA but did not observe any evidence for such associations (data not shown). Nevertheless, further investigations would be warranted to assess whether this original methylation signal was due to random fluctuations or could suggest unsuspected biological links between thrombin generation, *C15orf41* methylation and anaemia.

According to our previous work [14], our study was well powered to detect, at the statistical threshold of $1.29 * 10^{-7}$, any DNA methylation change explaining at least 6% of the variability of a quantitative trait. As a consequence, it may then be speculated that: 1- whole blood DNA may not be a good model for identifying DMRs associated with thrombin generation or 2- variability of DNA methylation levels measured in whole blood cells has weaker effects, if any, on plasma levels of thrombin generation, effects that would require a much larger sample size in order to be detected. In that perspective, the summary statistics derived from the meta-analysis of our two MWAS cohorts on TGP are available upon request for those who would like to combine their own results with ours. In addition, Illumina HumanMethylation450 array and TGP phenotype data from MARTHA participants used in this work are available in the ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-MTAB-3127. Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.thromres.2014.12.010.

Authorship Contributions

A. Rocañín-Arjó conducted the statistical analysis of genome-wide methylation data and drafted the short report. J. Dennis, P. Suchon, D. Aïssi and V. Truong collected, pre processed and analyzed data. PE. Morange, F. Gagnon and DA. Trégouët designed the research studies, coordinated the analyses and wrote the manuscript.

Conflict of Interest Statement

No conflict of interest to disclose for the study.

Acknowledgements

ARA was supported by a PhD grant from the Région Ile de France (CORDDIM). The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique. The F5L Thrombophilia French-Canadian Pedigree study was supported by grants from the Canadian Institutes of Health Research

(MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). The Human450Meethylationepityping was partially funded by the Canadian Institutes of Health Research (grant MOP 86466) and by the Heart and Stroke Foundation of Canada (grant T6484).FG holds a Canada Research Chair. Statistical analyses were performed using the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05).

Author details
[1] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, F-75013 Paris, France. [2] INSERM, UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, F-75013 Paris, France. [3] ICAN Institute for Cardiometabolism and Nutrition, F-75013 Paris, France. [4] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada. [5] Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France. [6] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, F-13385 Marseille, France. [7] Laboratory of Haematology, La Timone Hospital, F-13385 Marseille, France.

References

1. Peng, Y., Jahroudi, N.: The NFY transcription factor inhibits von willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases **278**(10), 8385–8394. doi:10.1074/jbc.M213156200. Accessed 2015-10-14
2. Friso, S., Lotto, V., Choi, S.-W., Girelli, D., Pinotti, M., Guarini, P., Udali, S., Pattini, P., Pizzolo, F., Martinelli, N., Corrocher, R., Bernardi, F., Olivieri, O.: Promoter methylation in coagulation f7 gene influences plasma FVII concentrations and relates to coronary artery disease **49**(3), 192–199. doi:10.1136/jmedgenet-2011-100195. Accessed 2015-10-14
3. El-Maarri, O., Becker, T., Junen, J., Manzoor, S.S., Diaz-Lacava, A., Schwaab, R., Wienker, T., Oldenburg, J.: Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males **122**(5), 505–514. doi:10.1007/s00439-007-0430-3. Accessed 2015-10-14
4. Murphy, T.M., Mill, J.: Epigenetics in health and disease: heralding the EWAS era **383**(9933), 1952–1954. doi:10.1016/S0140-6736(14)60269-5. Accessed 2015-10-14
5. Hemker, H.C., Giesen, P., Al Dieri, R., Regnault, V., de Smedt, E., Wagenvoord, R., Lecompte, T., B\éguin, S.: Calibrated automated thrombin generation measurement in clotting plasma **33**(1), 4–15. doi:10.1159/000071636. Accessed 2015-10-14
6. Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Kallel, C., Zelenika, D., Lathrop, M., Trégouët, D.-A., Morange, P.-E.: Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein c anticoagulant pathway: the MARTHA project **157**(2), 230–239. doi:10.1111/j.1365-2141.2011.09025.x. Accessed 2015-10-14
7. Dick, K.J., Nelson, C.P., Tsaprouni, L., Sandling, J.K., Aïssi, D., Wahl, S., Meduri, E., Morange, P.-E., Gagnon, F., Grallert, H., Waldenberger, M., Peters, A., Erdmann, J., Hengstenberg, C., Cambien, F., Goodall, A.H., Ouwehand, W.H., Schunkert, H., Thompson, J.R., Spector, T.D., Gieger, C., Trégouët, D.-A., Deloukas, P., Samani, N.J.: DNA methylation and body-mass index: a genome-wide analysis **383**(9933), 1990–1998. doi:10.1016/S0140-6736(13)62674-4. Accessed 2015-10-14
8. Rocanin-Arjo, A., Cohen, W., Carcaillon, L., Frère, C., Saut, N., Letenneur, L., Alhenc-Gelas, M., Dupuy, A.-M., Bertrand, M., Alessi, M.-C., Germain, M., Wild, P.S., Zeller, T., Cambien, F., Goodall, A.H., Amouyel, P., Scarabin, P.-Y., Trégouët, D.-A., Morange, P.-E., Consortium, a.t.C.: A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential **123**(5), 777–785. doi:10.1182/blood-2013-10-529628. Accessed 2015-10-14
9. Antoni, G., Morange, P.-E., Luo, Y., Saut, N., Burgos, G., Heath, S., Germain, M., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Zelenika, D., Alessi, M.-C., Drouet, L., Visvikis-Siest, S., Wells, P.S., Lathrop, M., Emmerich, J., Tregouet, D.-A., Gagnon, F.: A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism **8**(12), 2671–2679. doi:10.1111/j.1538-7836.2010.04092.x. Accessed 2015-10-14
10. Koestler, D.C., Christensen, B.C., Karagas, M.R., Marsit, C.J., Langevin, S.M., Kelsey, K.T., Wiencke, J.K., Houseman, E.A.: Blood-based profiles of DNA methylation predict the underlying distribution of cell types **8**(8), 816–826. doi:10.4161/epi.25430. Accessed 2015-10-14
11. Mägi, R., Morris, A.P.: GWAMA: software for genome-wide association meta-analysis **11**(1), 288. doi:10.1186/1471-2105-11-288. Accessed 2015-10-14
12. Tarantini, L., Bonzini, M., Tripodi, A., Angelici, L., Nordio, F., Cantone, L., Apostoli, P., Bertazzi, P.A., Baccarelli, A.A.: Blood hypomethylation of inflammatory genes mediates the effects of metal-rich airborne pollutants on blood coagulation **70**(6), 418–425. doi:10.1136/oemed-2012-101079. Accessed 2015-10-14
13. Babbs, C., Roberts, N.A., Sanchez-Pulido, L., McGowan, S.J., Ahmed, M.R., Brown, J.M., Sabry, M.A., Consortium, W., Bentley, D.R., McVean, G.A., Donnelly, P., Gileadi, O., Ponting, C.P., Higgs, D.R., Buckle, V.J.: Homozygous mutations in a predicted endonuclease are a novel cause of congenital dyserythropoietic anemia type i **98**(9), 1383–1387. doi:10.3324/haematol.2013.089490. Accessed 2015-10-14
14. Aïssi, D., Dennis, J., Ladouceur, M., Truong, V., Zwingerman, N., Rocanin-Arjo, A., Germain, M., Paton, T.A., Morange, P.-E., Gagnon, F., Trégouët, D.-A.: Genome-wide investigation of DNA methylation marks associated with FV leiden mutation **9**(9), 108087. doi:10.1371/journal.pone.0108087. Accessed 2015-10-14

4 Recherche de facteurs épigénétiques pouvant expliquer la pénétrance incomplète du Facteur V de Leiden

Un des premiers travaux que j'ai entrepris au cours de mon projet doctoral a été de déterminer si la mutation du Facteur 5 de Leiden (également appelée F5 R506Q ou rs6025 T/C) pouvait être associée aux niveaux de méthylation de certains sites CpG tout au long du génome afin d'identifier de nouveaux mécanismes pouvant expliquer la pénétrance incomplète de cette mutation sur le risque de thrombose veineuse (Aïssi *et al.* 2014).

La thrombose veineuse est une pathologie complexe commune caractérisée par une héritabilité estimée comprise entre 30% et 60% et un risque de développer la maladie chez une personne dont un germain a déjà développé la maladie multiplié par environ deux par rapport à la population générale. La mutation du Facteur 5 de Leiden est une variation génétique assez fréquente, environ 5% de la population, et associée à un risque d'environ 2,5 de développer la maladie à l'état hétérozygote. La pénétrance de cette mutation est incomplète avec seulement 10% des porteurs hétérozygotes et 80% des porteurs homozygotes qui développeront la maladie au cours de leur vie, avec une sévérité variable entre les individus touchés. Ces observations suggèrent que des facteurs génétiques et non génétiques contribuent à la pénétrance incomplète et l'hétérogénéité clinique observées.

L'objectif de mon travail était de déterminer si l'étude de la méthylation de l'ADN pouvait aider à identifier des marques épigénétiques pouvant expliquer une partie de la pénétrance incomplète de cette mutation. L'idée de ce travail était que l'identification de marques de méthylation différemment associés à la présence de la mutation pouvait mettre en lumière de nouveaux mécanismes épigénétiques liés à cette mutation. Jai donc réalisé une étude MWAS pour évaluer si le niveau de méthylation de sites CpG différait selon la présence ou non du polymorphisme chez 349 sujets de l'étude MARTHA (98 porteurs et 251 non-porteurs). L'analyse du niveau de méthylation de l'ADN, provenant de cellules du sang périphérique, de 388 120 sites CpG a permis d'identifier trois sites CpG du locus *SLC19A2* pour lesquels le niveau de méthylation différait significativement ($p < 3 * 10^{-8}$) entre les porteurs et les non porteurs du polymorphisme. Ces trois sites CpG ont été répliqués ($p < 2 * 10^{-7}$) dans l'étude indépendante F5L-Pedigrees composée de 214 individus répartis sur 5 grandes familles franco-canadiennes et dont 53 sont porteurs du polymorphisme contre 161 non-porteurs. Dans les deux études, ces trois sites CpG sont également associés ($2,33 * 10^{-11} < p < 3,02 * 10^{-4}$) avec des biomarqueurs de défaillance de la voie métabolique de la protéine C connus pour être influencés

par la mutation du Facteur V Leiden. Le gène *SLC19A2* étant situé à proximité du gène *F5*, j'ai entrepris une analyse plus détaillée de cette région.

Cette analyse a permis de mettre en évidence que l'association identifiée entre la mutation du Facteur V Leiden et les niveaux de méthylation au locus *SLC19A2* était due au déséquilibre de liaison entre la mutation du FV Leiden et un bloc de polymorphismes situés au locus *SLC19A2*. Après ajustement sur les polymorphismes de *SLC19A2*, dont rs970740, trouvés également fortement associés aux niveaux de méthylation des trois sites CpG de *SLC19A2*, la mutation du FV Leiden n'était plus associée ($p > 0,05$) aux niveaux de méthylation de *SLC19A2*. Cette étude, dont les résultats détaillés sont décrits dans la publication jointe à la fin de cette section, n'a donc pas permis de mettre en évidence des mécanismes épigénétiques pouvant expliquer la pénétrance incomplète de la mutation du Facteur V Leiden. En revanche, ce travail a permis de suggérer que la méthylation du gène *SLC19A2* mesurée dans l'ADN sanguin, pouvait être sous l'influence de polymorphismes situés dans ce gène. Le gène *SLC19A2* est impliqué dans l'anémie mégaloblastique thiamine-dépendante (Beshlawi *et al.* 2014 ; Diaz *et al.* 1999 ; G. Liu *et al.* 2014) mais les mécanismes sous-jacents sont mal connus (Mikstiene *et al.* 2015 ; Tahir *et al.* 2015). Les mécanismes de méthylation mis en évidence via ce travail pourraient permettre de mieux comprendre l'implication de *SLC19A2* dans cette pathologie.

OPEN ACCESS Freely available online

PLOS ONE

Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leiden Mutation

Dylan Aïssi[1,2,3], Jessica Dennis[4], Martin Ladouceur[4,5], Vinh Truong[4], Nora Zwingerman[4], Ares Rocanin-Arjo[1,2,3], Marine Germain[1,2,3], Tara A. Paton[6], Pierre-Emmanuel Morange[7,8,9], France Gagnon[4], David-Alexandre Trégouët[1,2,3]*

1 Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team *Genomics & Pathophysiology of Cardiovascular Diseases*, Paris, France, **2** INSERM, UMR_S 1166, Team *Genomics & Pathophysiology of Cardiovascular Diseases*, Paris, France, **3** ICAN Institute for Cardiometabolism and Nutrition, Paris, France, **4** Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Canada, **5** Centre de Recherches du CHUM, Montréal, Canada, **6** The Centre for Applied Genomics and Program in Genetics and Genome Biology, The Hospital for Sick Children, Toronto, Ontario, Canada, **7** Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Marseille, France, **8** INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Marseille, France, **9** Laboratory of Haematology, La Timone Hospital, Marseille, France

Abstract

In order to investigate whether DNA methylation marks could contribute to the incomplete penetrance of the FV Leiden mutation, a major genetic risk factor for venous thrombosis (VT), we measured genome-wide DNA methylation levels in peripheral blood samples of 98 VT patients carrying the mutation and 251 VT patients without the mutation using the dedicated Illumina HumanMethylation450 array. The genome-wide analysis of 388,120 CpG probes identified three sites mapping to the *SLC19A2* locus whose DNA methylation levels differed significantly ($p<3\ 10^{-8}$) between carriers and non-carriers. The three sites replicated ($p<2\ 10^{-7}$) in an independent sample of 214 individuals from five large families ascertained on VT and FV Leiden mutation among which 53 were carriers and 161 were non-carriers of the mutation. In both studies, these three CpG sites were also associated ($2.33\ 10^{-11}<p<3.02\ 10^{-4}$) with biomarkers of the Protein C pathway known to be influenced by the FV Leiden mutation. A comprehensive linkage disequilibrium (LD) analysis of the whole locus revealed that the original associations were due to LD between the FV Leiden mutation and a block of single nucleotide polymorphisms (SNP) located in *SLC19A2*. After adjusting for this block of SNPs, the FV Leiden mutation was no longer associated with any CpG site ($p>0.05$). In conclusion, our work clearly illustrates some promises and pitfalls of DNA methylation investigations on peripheral blood DNA in large epidemiological cohorts. DNA methylation levels at SLC19A2 are influenced by SNPs in LD with FV Leiden, but these DNA methylation marks do not explain the incomplete penetrance of the *FV* Leiden mutation.

Citation: Aïssi D, Dennis J, Ladouceur M, Truong V, Zwingerman N, et al. (2014) Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leiden Mutation. PLoS ONE 9(9): e108087. doi:10.1371/journal.pone.0108087

Editor: Tanja Zeller, Medical University Hamburg, University Heart Center, Germany

Received May 7, 2014; **Accepted** August 12, 2014; **Published** September 29, 2014

Data Availability: The authors confirm that, for approved reasons, some access restrictions apply to the data underlying the findings. Detailed results of the discovery MWAS analysis are available upon request by contacting Dr David-Alexandre Trégouët at david.tregouet@upmc.fr. Raw data cannot be made publicly available because consent agreement signed by patients does not allow such. However, specific data access can be requested by contacting Pr Pierre-Emmanuel Morange at pierre.morange@ap-hm.fr.

Funding: DA and ARA were supported by a PhD grant from the Région Ile de France (CORDDIM). JD holds a Vanier Canada Graduate Scholarship, ML holds a fellowship from the Canadian Institutes of Health Research (CIHR), and FG holds a Canada Research Chair. JD, ML and NZ are fellows of the CIHR Strategic Training for Advanced Genetic Epidemiology Training Grant (STAGE) in Genetic Epidemiology and Statistical Genetics (GET-101831). The MARTHA project was supported by a grant from the Program Hospitalier de Recherche Clinique. The H450M epityping was partially funded by the Canadian Institutes of Health Research (grant MOP 86466) and by the Heart and Stroke Foundation of Canada (grant T6484). Statistical analyses of the MARTHA datasets was performed using the C2BIG computing cluster funded by the Région Ile de France, Pierre and Marie Curie University, and the Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05). The F5L Thrombophilia French-Canadian Pedigree study was supported by grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: david.tregouet@upmc.fr

Introduction

Venous thrombosis (VT) is a common complex disease characterized by a sibling relative risk of ~3 [1] and heritability estimates ranging from 30% to 60% [2,3]. Contrary to other complex diseases, few new VT susceptibility genes were discovered by the recent waves of Genome-Wide Association Studies (GWAS) [4]. Established VT-associated genes collectively explain only about 5% of the disease heritability [2] and family history of VT remains an important risk factor despite adjustment for known variants [5,6]. In addition, marked clinical variability is observed even in affected individuals from the same family and carrying the same mutation [6]. In particular, the penetrance of the *FV* Leiden mutation (i.e. *F5* R506Q or rs6025T/C), one of the major VT genetic risk factors present in about 5% of the general population, is quite low, only 10% of heterozygotes and 80% of homozygotes develop VT in their lifetime, with varying severity among affected individuals. These observations suggest that additional genetic and non-genetic factors contribute to the incomplete penetrance of FV

Leiden and the clinical heterogeneity VT, as well as idiopathic VT.

Several lines of evidence support the role of DNA methylation marks as contributing factors in complex human diseases, including thrombosis-related disorders [7–11]. For example, quantitative risk factors for VT such as body-mass-index [12] and levels of von Willebrand factor [13], Factor VIII [14], and homocysteine [15] have been associated with DNA methylation marks. Further, lifestyle and environmental VT risk factors, such as smoking and air pollution, have been associated with methylation levels in genes relevant to VT pathophysiological mechanisms [16–18]. Until recently, such investigations were restricted to experimental models or small study samples, and restricted to candidate genomic regions.

The recent enthusiasm for agnostic investigations of methylation marks in peripheral blood DNA as a mean to investigate complex disease etiology and to generate novel mechanistic hypotheses is justified [19–21]. First,genome-wide methylation arrays, such as the Illumina HumanMethylation450 bead array, are now widely recognized as robust and efficient tools for epidemiological studies aiming at identifying methylation marks at CpG sites associated with environmental and genetic risk factors [12,22,23]. Second, biobanked peripheral blood DNA has been shown to be a robust and practical model for epidemiological epigenetic investigations [12,24–26]. Third, evidence of peripheral blood DNA methylation marks as surrogates for methylation marks at other disease relevant tissues and cell types are increasingly emerging [12,23,24]. As whole blood DNA methylation levels reflect the average level resulting from the epigenetic state at different cell types, the identification of DNA methylation marks in peripheral blood cells may point out to novel biological mechanisms that subsequently can be validated in the principal effector cell types where stronger associations are expected [12]. Finally, and specific to this study, DNA from peripheral blood originates mainly from leukocytes, which are key effector cells for both coagulation and inflammation, the two principal pathophysiological mechanisms underlying VT.

In the current work, we hypothesized that DNA methylation marks contribute to the incomplete penetrance of the FV Leiden mutation. We undertook a DNA methylome-wide association scan (MWAS, sometimes referred to as EWAS which stands for Epigenome-Wide Association Scan) to identify DNA methylation changes in relation to the presence/absence of the *F5* rs6025 mutation in 349 (98/251) MARTHAVT patients. Main findings were replicated in an independent study of 214 (53/161) individuals, processed with the same Illumina array.

Material and Methods

Ethics Statement

For MARTHA, ethics approval was obtained from the "Département santé de la direction générale de la recherche et de l'innovation du ministère" (Projects DC: 2008-880 & 09.576).

For the F5L-families study, ethics approval was obtained from the Ottawa Hospital and the University of Toronto ethics boards. All subjects in both studies provided written informed consent in accordance with the Declaration of Helsinki.

Study populations

Discovery study sample. The MARTHA study is a collection of 1,542 patients with VT recruited from the Thrombophilia centre of La Timone hospital (Marseille, France) [27–30]. All subjects had a documented history of VT, were free of chronic conditions, and were free of inherited thrombophilia including: anti-thrombin, protein C and protein S deficiencies and homozygosity for the Factor V Leiden and Factor II G20210A mutations. For the current project, 349 MARTHA patients were randomly selected for DNA methylation analysis.

Replication study sample. The family study is composed of five extended French-Canadian pedigrees, totaling 255 relatives, ascertained at the Thrombosis Clinic of the Ottawa Hospital through single probands with idiopathic VT and heterozygote for the Factor V Leiden mutation. Probands were free of acquired VT risk factors such as cancer, myeloproliferative disease, pregnancy, puerperium, prolonged immobilization, trauma, surgery and antiphospholipid syndrome, and were free of inherited thrombophilia (see above). A detailed description of this study can be found in [27]. Only 218 family members for whom DNA was still available were included in the current work.

Genome wide DNA methylation assay

Genomic DNA was isolated from peripheral blood cells using an adaptation of the method proposed by [31]. For each sample, 1 µg genomic DNA was bisulphite converted using the Qiagen EpiTect 96 Bisulfite Kit. Then, 200 ng of bisulfite-converted DNA at 50 ng/µl was independently amplified, labeled, and hybridized to Infinium HumanMethylation450 BeadChip microarrays [25] and scanned with default settings using the Illumina iScan. This Illumina array covers 99% of RefSeq genes and surveys the DNA methylation levels at 482,421 CpG sites, with an average of 17 CpG sites per gene region. The discovery and replication samples were processed simultaneously at The Center for Applied Genomics (TCAG, Toronto, Canada).

Quality controls and normalization procedures

From the 485,577 probes available on the Illumina array, we excluded from further analyses probes that measured single nucleotide polymorphisms (n = 65), that are either cross-reactive (n = 30,969) or polymorphic at the targeted CpG site (n = 66,877) [32,33]. Of note, 4,464 probes shared the two last features.

DNA methylation data were expressed as a β-value, a continuous variable over the [0–1] interval, representing the percentage of methylation of a given CpG site [34]. Methylation values were corrected for background by use of the Noob method implemented in the "methylumi" package [35], for dye bias following the manufacturer's recommendation (http://support.illumina.com/downloads/genomestudio_m_module_v18_ug_%2811319130_b%29.ilmn) and normalized for design type bias according to the SWAN method [36] implemented in the minfi R package [37].

Quality control and normalization were done simultaneously on the MARTHA and F5L-families datasets. Probes (n = 4,010) with a detection p-value (as described in the "minfi" package) greater than 0.05 in more than 5% of the total processed samples were then excluded from further analyses. Principal components analysis was carried out on probe data to detect outliers and four F5L-families individuals were then excluded. This led to a final selection of 388,120 probes (among which 1,289 tagged for CpH sites) that were tested for association with the presence/absence of the *FV* Leiden mutation-tagging rs6025-C allele.

Biological measurements

In MARTHA, we used the Agkistrodon contortrix venom (ACV) test as a quantitative biomarker of the protein C pathway. The ACV test was expressed as a normalized ACV value (ACVn) as described in [30]. The ACVn ratio was available in 260 MARTHA patients with DNA methylation measurements. A complete blood count, including white blood cell types (neutrophils, lymphocytes, monocytes, eosinophils, and basophils), was

Table 1. Characteristics of the studied populations.

	MARTHA	F5L-famiiles
	N = 349	N = 214
Mean age in yrs ± SD	43.8±14.1	39.6±16.7
Males/Females	75/274	101/113
VT patients (%)	349 (100%)	11 (5.1%)
Heterozygote carriers of the *F5* rs6025	98	53
ACVn ratio[(1)]	0.89±0.38	NA
F5 rs6025 carriers	0.52±0.10	NA
Non-carriers	1.09±0.32	NA
APCR ratio [(2)]	NA	2.56±0.67
F5 rs6025 carriers	NA	1.67±0.19
Non-carriers	NA	2.86±0.49

[(1)] In MARTHA, ACVn ratio was significantly ($p = 1.63\ 10^{-38}$) decreased in *F5* rs6025 carriers compared to non-carriers.
[(2)] In families, APCR ratio was significantly ($p = 9.98\ 10^{-47}$) decreased in *F5* rs6205 carriers compared to non-carriers.
doi:10.1371/journal.pone.0108087.t001

determined by ADVIA 120 Hematology System (Siemens Healthcare Diagnostics, Deerfield, IL).

In F5L-families, activated protein C resistance (APCR) levels were determined in 208 individuals using the APC-aPTT assay. The results of the test are expressed as the APC-sensitivity ratio, which is the quotient of the activated partial thromboplastin time (aPTT) of the plasma sample with and without exogenous APC [38].

Genotyping

MARTHA patients were genotyped with the Illumina Human 610/660W-Quad beadchips [28]. Autosomal SNPs that satisfied quality control criteria (n = 481,002) [39] were then used for imputing SNPs from the 1000 Genomes 2012-02-14 release reference dataset. Imputation was performed by use of MACH (v1.0.18.c) software [40]. All SNPs with acceptable imputation quality ($r^2 > 0.3$) [41], minor allele frequency>0.01 and mapping to the chromosome 1 169101258–169555769 locus were tested for association with *SLC19A2* probes.

The F5L-families study was genotyped with the Illumina 660W-Quad beadchip. Detailed description of the quality control procedure is available in [28].

Statistical Analysis

Discovery MWAS. Because methylation β-values are often not normally distributed, exhibiting bi-modality, right-, or left-tailed skewed distributions, our discovery MWAS was performed using a logistic regression model with carrier status (yes or no) as the outcome and the β-values as covariates. Any methylation probes that satisfied the Bonferroni threshold of 1.29×10^{-7} (~0.05/388,120) were selected and their distribution was assessed (Figure S1). For uni-modal probes, a linear regression model was also applied with β-values as the outcome and carrier status as the covariate to assess the consistency of the MWAS results and to provide an estimate of the effect of the rs6025 variant on the DNA

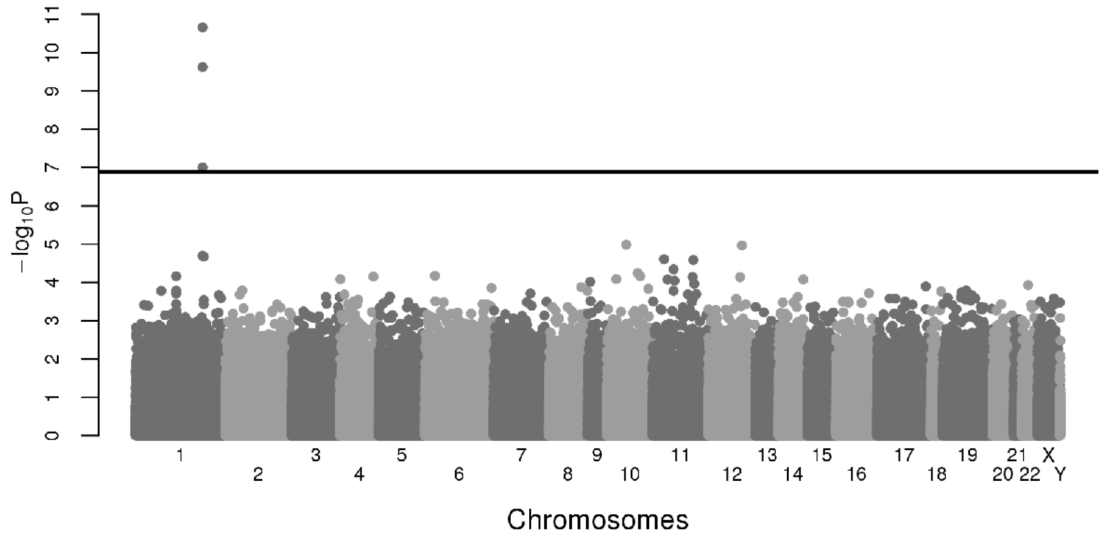

Figure 1. Manhattan plot of the MWAS results at 388,120 CpG sites.
doi:10.1371/journal.pone.0108087.g001

Table 2. Association[1] of *SLC19A2*CpG sites with rs6025 (FV Leiden mutation) in the discovery and replication studies.

	Discovery MARTHA study			Replication F5L-families study		
	Non-Carriers (N = 251)	Carriers (N = 98)	Association Test p-value[2]	Non-Carriers (N = 161)	Carriers (N = 53)	Association Test p-value[2]
cg16548605	0.93 (0.02)	0.89 (0.03)	$1.90\ 10^{-29}$	0.93 (0.02)	0.90 (0.03)	$6.58\ 10^{-14}$
cg04083076	0.73 (0.06)	0.64 (0.06)	$5.73\ 10^{-22}$	0.74 (0.06)	0.67 (0.09)	$1.19\ 10^{-10}$
cg09671955	0.53 (0.06)	0.48 (0.07)	$3.49\ 10^{-12}$	0.55 (0.07)	0.50 (0.07)	$5.62\ 10^{-7}$

[1] Association is expressed as methylation β-value mean (SE) in carriers and non-carriers.
[2] Reported p-values were those derived from a linear regression model where the probe methylation level was the outcome and the carrier status the covariate, while adjusting for age, sex, batch, chip and cell type composition.
doi:10.1371/journal.pone.0108087.t002

methylation β-value. A linear regression model with M-transformed values [42,43] instead of β-values as outcome was also applied to provide further statistical support to the obtained results (Table S1). Analyses were adjusted for age, sex, batch and chip effects [44]. Because DNA methylation levels measured in peripheral blood DNA reflect the average level of DNA methylation in different cell types including lymphocytes, neutrophils, basophils and eosinophils, all analyses were also adjusted for cell type composition to avoid any contamination bias [45–47]. For this, we used specific biological counts of lymphocytes, monocytes, neutrophils, eosinophils and basophils available for all MARTHA samples to characterize cell type composition.

Replication study. In the F5L-families study, association of selected probes with rs6025 was assessed using the linear model mentioned above, after having checked for the uni-modality of the data distribution (Figure S1). In order to handle correlations between family data, a linear mixed regression model as implemented in the NMLE R package (http://cran.r-project.org/web/packages/nlme/) was employed where the family variable was defined as a random effect. Analyses were adjusted for age, sex, batch, chip and cell type composition. As specific cell type counts were not available in the family study, adjustment for cell type composition was handled by the method described in [48,49]. The Bonferroni corrected threshold of 0.0167 (= 0.05/3) was used for declaring replication.

Further analyses. Association of selected probes with quantitative biomarkers was tested using a linear (mixed in F5L-families) model where log-transformed biomarker values were used as the outcome and the methylation β-values as covariates. Models were adjusted by the same covariates as described above.

The association of imputed SNPs with methylation β-values was tested by entering the allele dosage of the imputed SNP as a covariate in a linear regression model with β-values as the outcome. The allele dosage is a real number ranging from 0 to 2 corresponding to the expected number of minor alleles computed from the posterior probabilities of possible imputed genotypes.

To get more power for detecting CpG sites variability associated with the *F5* rs6025, we finally performed a combined analysis of both MARTHA and F5L-families studies. For this part, linear regression analyses (mixed linear model in F5L-families) were conducted for each CpG β-value with the rs6025 as covariate while adjusting for the same variables as indicated above. Regression coefficients associated with the rs6025 were then combined into a random-effect meta-analysis using the GWAMA program [50].

Results

Brief characteristics of the two studied populations are given in Table 1. To support the validity of the discovery MARTHA DNA methylation dataset, we investigated two previously reported robust associations with DNA methylation marks, the association of smoking with decreased methylation levels at *F2RL3* CpG cg03636183 [16,22] and the association of rs713586 with DNAJC27/ADCY3 CpG cg01884057 [51]. Consistent and strong significant associations were observed in MARTHA. Current smokers exhibited lower levels of methylation at cg03636183 compared to non-smokers and former smokers ($p = 1.13\times10^{-29}$) (Figure S2). The rs713586-T allele was associated with decreased methylation β-values at cg01884057 in a fairly additive manner ($p = 7.38\times10^{-68}$) (Figure S3).

A Manhattan plot of the MWAS results is shown in Figure 1. Three CpG sites, all mapping the *SLC19A2* gene region, were associated with rs6025 at the genome-wide significant threshold of 1.29×10^{-7} (~0.05/388,120). DNA methylation levels at these sites, cg16548605 ($p = 3.61\times10^{-11}$), cg04083076 ($p = 2.82\times10^{-10}$) and cg09671955 ($p = 2.66\times10^{-8}$), were decreased in carriers of the rs6025-C allele compared to non-carriers (Table 2, Figure S4).

Table 3. Association[1] of *SLC19A2* CpG sites with ACVn (MARTHA) and APCR (F5L-families) phenotypes.

	MARTHA study (N = 260)		F5L-families study (N = 208)	
	raw[2]	Adjusted for rs6025	raw[2]	Adjusted for rs6025
cg16548605	46.9 (32.6–61.1) $p = 8.14\ 10^{-10}$	−1.2 (−14–11.7) p = 0.86	58.1 (44–72.1) $p = 1.11\ 10^{-13}$	13.6 (3.1–24.1) p = 0.01
cg04083076	18.1 (11.5–24.7) $p = 2.12\ 10^{-7}$	−3.1 (−8.8–2.50) p = 0.28	21.9 (15.1–28.7) $p = 2.13\ 10^{-9}$	3.9 (−0.8–8.5) p = 0.10
cg09671955	15.1 (7.4–22.9) $p = 1.7\ 10^{-4}$	−1.9 (−7.8–4.0) p = 0.53	14.4 (7.3–21.4) $p = 1.02\ 10^{-4}$	−1.3 (−4.6–4.1) p = 0.90

[1] Association is expressed as % change in phenotype (95% Confidence Interval) for every 0.1 unit increase in methylation β-value.
[2] Analysis were adjusted for age, sex, batch, chip and cell type composition.
doi:10.1371/journal.pone.0108087.t003

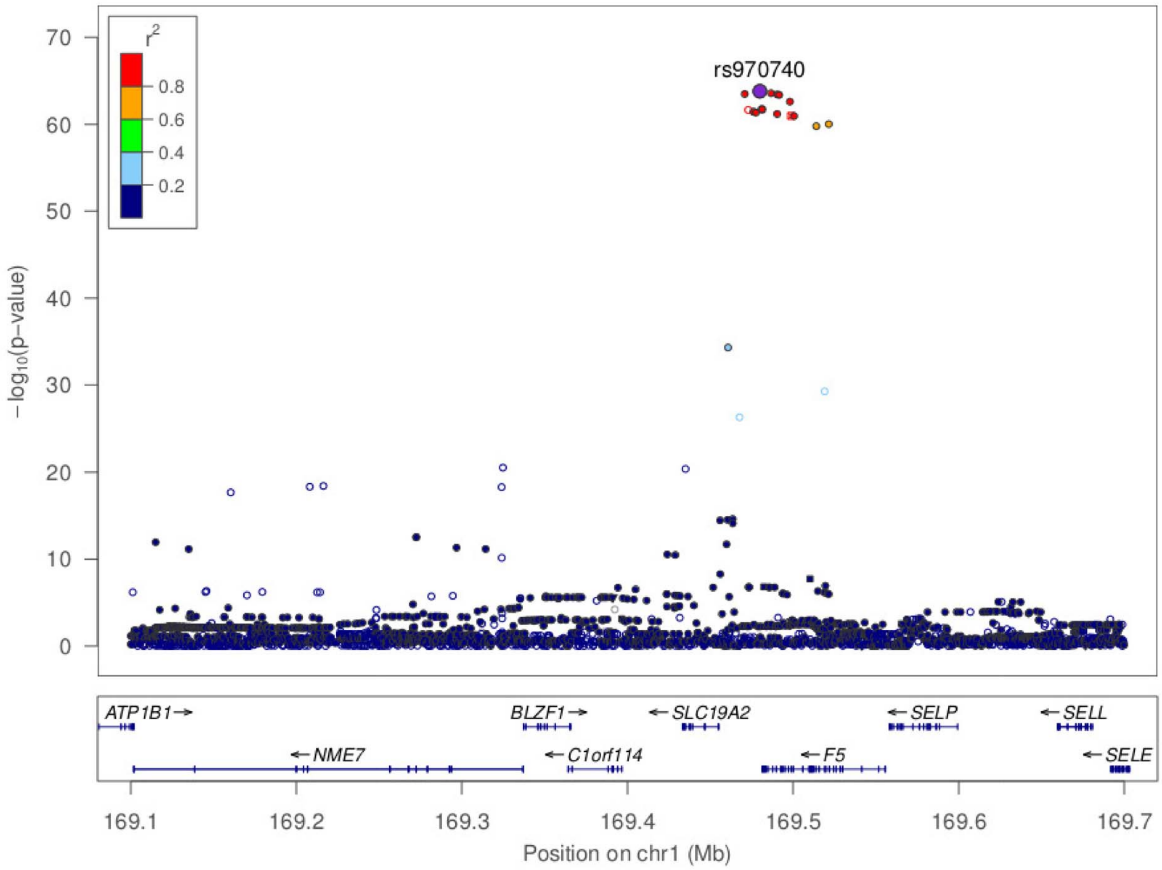

Figure 2. Region Association plot of the association between chromosome 1q23.3 SNPs and cg16548605 CpG site variability in the MARTHA study.
doi:10.1371/journal.pone.0108087.g002

Of note, the strongest association observed with any *F5* CpG site was p = 0.045 for cg16054275.

These three CpG probes were tested for replication in 214 related individuals from the F5L Thrombophilia French-Canadian Pedigree study [27,28], referred thereafter to as the F5L-families study. In this independent study, all three *SLC19A2* probes also exhibited lower DNA methylation levels in carriers (n = 53) compared to in non-carriers (n = 161) of the rs6025-C allele (Table 2; Figure S5).

To further validate these results, using a linear model, we assessed the association of the 3 *SLC19A2* probes with quantitative biomarkers of the Protein C pathway known to be under the strong influence of rs6025: the Agkistrodon contortrix venom test (ACVn) [30,52] in the discovery MARTHA population and the activated protein C resistance (APCR) test [53] in the replication family study. In both studies, these biomarkers demonstrated decreased levels in carriers of the *F5* rs6025-C allele compared to non-carriers (Table 1). The three *SLC19A2* CpG sites were significantly associated with the two biomarkers, with all p-values$<10^{-3}$ (Table 3). For example, every 0.1 unit increase in the methylation β-value at cg16548605 was associated with a 46.9% (95% confidence interval: 32.6%–61.1%) higher ACVn value in the MARTHA population and with a 50.0% (95%CI: 36.3%–63.7%) higher APCR value in the F5L-families. After adjusting for rs6025, these associations completely vanished, with all p-values >0.01 (Table 3).

Because the *SLC19A2* gene maps to chromosome 1q23.3 in the vicinity of the *F5* gene, a locus known to exhibit strong linkage disequilibrium (LD) over a large genomic distance of ~460 Kb [39] (Figure S6), one cannot rule out the possibility that the associations between rs6025 and methylation at *SLC19A2* probes were due to other SNPs in LD with rs6025. We therefore examined the association of the methylation levels of the three *SLC19A2* probes with 3,213 SNPs at this locus using genome-wide SNP data available in the MARTHA study. Results of these association analyses, where DNA methylation levels were the outcome and the SNPs the predictors, are illustrated in Figure 2. The strongest association for cg16548605 was observed with rs970740 ($p = 1.61\times10^{-66}$) where the minor C allele was associated with decreased cg16548605 methylation levels (Table 4) (regression coefficient for adjusted allele effect $\beta = -0.049 \pm 0.0022$). The same pattern of associations was observed in the F5L-families study (Table 4). The C allele was also associated with decreased ACVn values ($\beta = -0.415 \pm 0.043$, $p = 3.20\times10^{-18}$) (Table 5). Interestingly, in a joint model where both rs970740 and rs6025 were used as covariates for predicting cg16548605 variability, the effect of rs970740 was highly significant ($p = 1.05\times10^{-38}$) but that of rs6025 was not (p = 0.90). Conversely, in a similar joint model for predicting ACVn levels, only the effect of rs6025 was significant ($p = 1.65\times10^{-20}$) while the effect of rs970740 completely vanished (p = 0.79). Rs970740 lies in the upstream *SLC19A2*/downstream *F5* region and is in moderate LD ($r^2 = 0.65$) with the *F5* rs6025. Similar patterns were observed

Table 4. Association of rs970740 with *SLC19A2* cg1658605, cg0483076 and cg09671955 levels.

	MARTHA				F5L-families study			
	TT (N = 232)	TC (N = 114)	CC (N = 3)	P value	TT (N = 140)	TC (N = 68)	CC (N = 3)	P value
cg16548605	0.935 (0.010)	0.889 (0.027)	0.797 (0.045)	1.66×10^{-66}	0.930 (0.010)	0.900 (0.029)	0.845 (0.034)	4.49×10^{-33}
cg04083076	0.735 (0.051)	0.642 (0.056)	0.541 (0.135)	1.16×10^{-34}	0.752 (0.055)	0.681 (0.077)	0.607 (0.142)	1.65×10^{-20}
cg09671955	0.538 (0.059)	0.479 (0.066)	0.403 (0.063)	8.00×10^{-17}	0.556 (0.066)	0.503 (0.071)	0.490 (0.053)	2.76×10^{-10}

p-values were adjusted for age, sex, batch, chip and cell type composition.
In the F5L-families study, the rs970740 was not genotyped but substituted by the proxy rs2420371 that is in complete association ($r^2 = 1$) with it.
doi:10.1371/journal.pone.0108087.t004

for the other *SLC19A2* cg04083076 and cg09671955 CpG sites (data not shown).

These results demonstrate that two independent phenomena act at this locus: an effect of rs970740 (or its proxies) on the variability of *SLC19A2* methylation levels and the effect of *F5* rs6025 on the ACVn biomarker. The presence of LD between rs970740 and the *F5* rs6025 mutation confounds the associations between methylation at *SLC19A2* sites and both the *F5* rs6025 and the ACVn biomarker.

To improve statistical power and increase opportunity to detect smaller effect sizes of additional CpG sites and *F5* rs6025 associations, we combined the discovery and replication study samples into a meta-analysis. An additional CpG probe (cg26009832) mapping the *SLC19A2/F5* locus reached genome-wide significance ($p = 1.42\ 10^{-8}$).

Discussion

The starting hypothesis of this work was that DNA methylation marks associate with the *F5* rs6025 mutation and contribute to the incomplete penetrance of this strong genetic risk factor for VT. Thus, we undertook the first MWAS of the *F5* rs6025 in a large sample of 349 individuals and replicated the findings in an independent sample of 214 related subjects. We identified and replicated three CpG sites exhibiting a genome-wide significant difference in methylation levels in carriers and non-carriers of the mutation. These CpG sites were also strongly associated with the plasma variability of quantitative biomarkers influenced by the *F5* rs6025. However, when we integrated our MWAS and GWAS data, the observed associations between methylation levels at three CpG sites in *SLC19A2* and *F5* rs6025 were in fact due to LD between the rs6025 and SNPs located in *SLC19A2*.

We observed strong statistical evidence for association between the *SLC19A2* promoter rs970740 SNP (or any SNP in strong LD with it) and three identified *SLC19A2* CpG sites, independently of *F5* rs6025. According to public database, including 1000 Genomes, none of the probes measuring these three CpG sites are polymorphic and the rs970740 does not map to a CpG island. This strongly suggests the existence of variant(s) influencing the variability of DNA methylation levels at the *SLC19A2* gene. How the rs970740 T/C genetic variation (or any linked SNP) affects *SLC19A2* DNA methylation remains an open question. This could be through the creation of a transcription factor binding site, the modification of the local CpG sites distribution, or more complex phenomena [54–59]. The *SLC19A2* gene codes for a thiamine transporter protein that has been associated with human anemia syndrome [60–62]. Our results suggest that genetically determined DNA methylation sensitive mechanisms are involved in this disease susceptibility.

Several conclusions could be drawn from this work. First, three identified CpG sites were found to be strongly associated with the plasma variability of two quantitative biomarkers of the coagulation cascade, supporting the potential of genome-wide DNA methylation data to identify epigenetic marks associated with biological phenotypes involved in thrombotic disorders. Nonetheless, this works highlights the need for careful analyses of associations between genetic variants, biological phenotypes, and methylation at CpG sites to avoid false inference on functional variant(s), in particular due to LD extending over large genomic distances. Integrating MWAS, GWAS and biological data from the same individuals, as illustrated here, is key to elucidating these relationships. Second, if such cautions are taken, DNA methylation data can help to dissect the functional mechanisms associated with known disease-causing SNPs.

Table 5. Association of rs970740 and rs6025 with *SLC19A2* cg16548605 CpG and ACVn levels in the MARTHA study.

	cg16548605	ACVn (log)
Univariate analysis [(1)]		
rs970740	$\beta = -0.049$ (0.0022) $p = 1.61\ 10^{-66}$	$\beta = -0.415$ (0.043) $p = 3.20\ 10^{-18}$
rs6025	$\beta = -0.044$ (0.0035) $p = 1.9\ 10^{-29}$	$\beta = -0.653$ (0.042) $p = 3.58\ 10^{-37}$
Joint analysis [(2)]		
rs970740	$\beta = -0.050$ (0.0033) $p = 1.05\ 10^{-38}$	$\beta = -0.014$ (0.052) $p = 0.79$
rs6025	$\beta = 0.001$ (0.004) $p = 0.90$	$\beta = -0.641$ (0.062) $p = 1.65\ 10^{-20}$

Association is expressed as the additive effect of the minor alleles on the variability of cg16548605 and log ACVn (95% Confidence Interval) adjusted for age, sex, batch, chip and cell type composition. In the univariate analysis[(1)], one SNP at a time is used as a covariate for predicting the phenotype. In the joint analysis [(2)], both SNPs are simultaneously introduced as predictors in the linear regression models.

doi:10.1371/journal.pone.0108087.t005

Several limitations must be acknowledged. First, the design of our study may not be optimal. As we did not have access to a case-control study for VT with genome-wide DNA methylation data, we adopted a 'case-only' approach for our discovery stage. Such approach has been shown to be a valid alternative to detect gene × environment or gene × gene interactions [63,64]. We here used this strategy with the aim of identifying epigenetic factors that interact with the FV Leiden mutation to modulate the risk of VT. Since, in our replication study, 45 carriers were VT patients and the remaining 8 carriers were healthy individuals, we also looked into this dataset for specific methylation patterns associated with VT but the low sample size precludes from identifying any significant association (data not shown).

Second, because homozygosity for *FV* Leiden mutation was an exclusion criteria for the MARTHA study and no homozygote was observed in the F5L-families, our analysis only included heterozygous carriers which may have reduced our power to identify CpG sites under the strong influence of the mutation.

Third, while extremely dense, the used Illumina array does not cover all sites of the genome that could be subject to DNA methylation, we cannot exclude that some relevant methylation association has been missed.

Fourth, the sample size of our discovery study was large enough to detect, at the genome-wide level of $1.29\ 10^{-7}$, a 0.05 increase in the methylation β-value. Whether an increase of smaller magnitude in DNA methylation marks detected in whole blood could be biologically relevant remains an open question. Whole blood DNA methylation levels reflect the average levels resulting from the epigenetic state at different white blood cells. Therefore, the cell subtype and tissue specific methylation marks would show a weaker effect in whole blood compared to levels that could be measured in thrombosis-relevant effector cells (e.g. monocytes, endothelial cells, hepatocytes). This phenomenon was recently observed and discussed for other cardiovascular-related phenotypes [12,65]. Thus, we cannot rule out the possibility that a stronger influence of *F5* rs6025on DNA methylation levels exists in specific cell types or tissues, such as the liver where F5 is mainly synthetized.

Last, we observed evidence that *SLC19A2*, with genetically determined DNA methylation levels, is a methylation quantitative trait locus (mQTL). However, as we did not have access to gene expression data, we were not able to assess whether the observed genetic influence on *SLC19A2* DNA methylation levels is followed by a direct impact on *SLC19A2* expression. Further epi-mapping at this locus would be of great interest.

In conclusion, our work does not support the existence of DNA methylation marks that could explain the incomplete penetrance of the *F5* rs6025. The incomplete penetrance could be the result of complex haplotype effects at the *F5* locus, or interaction at this locus with other genetic or environmental exposures; such investigations would require alternative study designs and much larger sample sizes to detect effects.

This work does, however, illustrate the promises and pitfalls of MWAS on peripheral blood DNA in large epidemiological studies, and suggests that the anemia-associated *SLC19A2* gene is a mQTL.

Supporting Information

Figure S1 Density distributions of *SLC19A2* methylation probes in the MARTHA and F5L-families studies.
(PDF)

Figure S2 Association of smoking with methylation β-values at *F2RL3* CpG cg03636183 in the MARTHA study.
(PDF)

Figure S3 Association of rs713586 with methylation β-values at CpG site cg01884057 in the MARTHA study.
(PDF)

Figure S4 Boxplot of the association between the *F5* rs6025-C allele with *SLC19A2* methylation probes in the discovery MARTHA study.
(PDF)

Figure S5 Boxplot of the association between the *F5* rs6025-C allele with *SLC19A2* methylation probes in the replication F5L-families study.
(PDF)

Figure S6 Linkage Disequilibrium plot at the 1q23.3 locus in the MARTHA study. This plot was drawn with the Haploview software (*Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005 [PubMed ID: 15297300]*).
(PDF)

Table S1 Consistency between the statistical p-values derived from the linear analyses of β and M-transformed values.
(DOCX)

Author Contributions

Conceived and designed the experiments: PEM FG DAT. Performed the experiments: TAP. Analyzed the data: DA JD ML VT NZ ARA MG. Contributed reagents/materials/analysis tools: PEM FG DAT. Wrote the paper: DA JD ML PEM FG DAT.

References

1. Sørensen HT, Riis AH, Diaz LJ, Andersen EW, Baron JA, et al. (2011) Familial risk of venous thromboembolism: a nationwide cohort study. J Thromb Haemost JTH 9: 320–324. doi:10.1111/j.1538-7836.2010.04129.x.
2. Germain M, Saut N, Greliche N, Dina C, Lambert J-C, et al. (2011) Genetics of venous thrombosis: insights from a new genome wide association study. PloS One 6: e25581. doi:10.1371/journal.pone.0025581.
3. Souto JC, Almasy L, Borrell M, Blanco-Vaca F, Mateo J, et al. (2000) Genetic susceptibility to thrombosis and its relationship to physiological risk factors: the GAIT study. Genetic Analysis of Idiopathic Thrombophilia. Am J Hum Genet 67: 1452–1459.
4. Morange PE, Tregouet DA (2011) Lessons from genome-wide association studies in venous thrombosis. J Thromb Haemost JTH 9 Suppl 1: 258–264. doi:10.1111/j.1538-7836.2011.04311.x.
5. Bezemer ID, van der Meer FJM, Eikenboom JCJ, Rosendaal FR, Doggen CJM (2009) The value of family history as a risk indicator for venous thrombosis. Arch Intern Med 169: 610–615. doi:10.1001/archinternmed.2008.589.
6. Cohen W, Castelli C, Suchon P, Bouvet S, Aillaud MF, et al. (2014) Risk assessment of venous thrombosis in families with known hereditary thrombophilia: the MARseilles-NImes prediction model. J Thromb Haemost 12: 138–146. doi:10.1111/jth.12461.
7. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, et al. (2010) Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. PloS One 5: e14040. doi:10.1371/journal.pone.0014040.
8. Gao S, Skeldal S, Krogdahl A, Sørensen JA, Andreasen PA (2005) CpG methylation of the PAI-1 gene 5′-flanking region is inversely correlated with PAI-1 mRNA levels in human cell lines. Thromb Haemost 94: 651–660.
9. Ordovás JM, Smith CE (2010) Epigenetics and cardiovascular disease. Nat Rev Cardiol 7: 510–519. doi:10.1038/nrcardio.2010.104.
10. Wierda RJ, Geutskens SB, Jukema JW, Quax PHA, van den Elsen PJ (2010) Epigenetics in atherosclerosis and inflammation. J Cell Mol Med 14: 1225–1240. doi:10.1111/j.1582-4934.2010.01022.x.
11. Zhuang J, Peng W, Li H, Wang W, Wei Y, et al. (2012) Methylation of p15INK4b and expression of ANRIL on chromosome 9p21 are associated with coronary artery disease. PloS One 7: e47193. doi:10.1371/journal.pone.0047193.
12. Dick KJ, Nelson CP, Tsaprouni L, Sandling JK, Aïssi D, et al. (2014) DNA methylation and body-mass index: a genome-wide analysis. Lancet. doi:10.1016/S0140-6736(13)62674-4.
13. Peng Y, Jahroudi N (2003) The NFY transcription factor inhibits von Willebrand factor promoter activation in non-endothelial cells through recruitment of histone deacetylases. J Biol Chem 278: 8385–8394. doi:10.1074/jbc.M213156200.
14. El-Maarri O, Becker T, Junen J, Manzoor SS, Diaz-Lacava A, et al. (2007) Gender specific differences in levels of DNA methylation at selected loci from human total blood: a tendency toward higher methylation levels in males. Hum Genet 122: 505–514. doi:10.1007/s00439-007-0430-3.
15. Friso S, Choi S-W, Girelli D, Mason JB, Dolnikowski GG, et al. (2002) A common mutation in the 5,10-methylenetetrahydrofolate reductase gene affects genomic DNA methylation through an interaction with folate status. Proc Natl Acad Sci U S A 99: 5606–5611. doi:10.1073/pnas.062066299.
16. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H (2011) Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet 88: 450–457. doi:10.1016/j.ajhg.2011.03.003.
17. Bind M-A, Baccarelli A, Zanobetti A, Tarantini L, Suh H, et al. (2012) Air pollution and markers of coagulation, inflammation, and endothelial function: associations and epigene-environment interactions in an elderly cohort. Epidemiol Camb Mass 23: 332–340. doi:10.1097/EDE.0b013e31824523f0.
18. Tarantini L, Bonzini M, Tripodi A, Angelici L, Nordio F, et al. (2013) Blood hypomethylation of inflammatory genes mediates the effects of metal-rich airborne pollutants on blood coagulation. Occup Environ Med 70: 418–425. doi:10.1136/oemed-2012-101079.
19. Murphy TM, Mill J (2014) Epigenetics in health and disease: heralding the EWAS era. Lancet. doi:10.1016/S0140-6736(14)60269-5.
20. Callaway E (2014) Epigenomics starts to make its mark. Nature 508: 22. doi:10.1038/508022a.
21. Osório J (2014) Obesity: Looking at the epigenetic link between obesity and its consequences-the promise of EWAS. Nat Rev Endocrinol 10: 249. doi:10.1038/nrendo.2014.42.
22. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, et al. (2013) Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PloS One 8: e63812. doi:10.1371/journal.pone.0063812.
23. Frazier-Wood AC, Aslibekyan S, Absher DM, Hopkins PH, Sha J, et al. (2014) Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. J Lipid Res. doi:10.1194/jlr.M048504.
24. Bell JT, Tsai P-C, Yang T-P, Pidsley R, Nisbet J, et al. (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 8: e1002629. doi:10.1371/journal.pgen.1002629.
25. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, et al. (2011) Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics Off J DNA Methylation Soc 6: 692–702.
26. Terry MB, Delgado-Cruzata L, Vin-Raviv N, Wu HC, Santella RM (2011) DNA methylation in white blood cells: association with risk factors in epidemiologic studies. Epigenetics Off J DNA Methylation Soc 6: 828–837.
27. Antoni G, Morange P-E, Luo Y, Saut N, Burgos G, et al. (2010) A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. J Thromb Haemost JTH 8: 2671–2679. doi:10.1111/j.1538-7836.2010.04092.x.
28. Antoni G, Oudot-Mellakh T, Dimitromanolakis A, Germain M, Cohen W, et al. (2011) Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. BMC Med Genet 12: 102. doi:10.1186/1471-2350-12-102.
29. Huang J, Sabater-Lleal M, Asselbergs FW, Tregouet D, Shin S-Y, et al. (2012) Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. Blood 120: 4873–4881. doi:10.1182/blood-2012-06-436188.
30. Oudot-Mellakh T, Cohen W, Germain M, Saut N, Kallel C, et al. (2012) Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway: the MARTHA project. Br J Haematol 157: 230–239. doi:10.1111/j.1365-2141.2011.09025.x.
31. Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res 16: 1215.
32. Chen Y, Choufani S, Grafodatskaya D, Butcher DT, Ferreira JC, et al. (2012) Cross-reactive DNA microarray probes lead to false discovery of autosomal sex-associated DNA methylation. Am J Hum Genet 91: 762–764. doi:10.1016/j.ajhg.2012.06.020.
33. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, et al. (2013) Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics Off J DNA Methylation Soc 8: 203–209. doi:10.4161/epi.23470.
34. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, et al. (2011) High density DNA methylation array with single CpG site resolution. Genomics 98: 288–295. doi:10.1016/j.ygeno.2011.07.007.
35. Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD (2013) Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res 41: e90. doi:10.1093/nar/gkt090.
36. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13: R44. doi:10.1186/gb-2012-13-6-r44.
37. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, et al. (2014) Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays. Bioinforma Oxf Engl. doi:10.1093/bioinformatics/btu049.
38. Dahlbäck B (1994) Inherited resistance to activated protein C, a major cause of venous thrombosis, is due to a mutation in the factor V gene. Haemostasis 24: 139–151.
39. Germain M, Saut N, Oudot-Mellakh T, Letenneur L, Dupuy A-M, et al. (2012) Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. PloS One 7: e38538. doi:10.1371/journal.pone.0038538.
40. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol 34: 816–834. doi:10.1002/gepi.20533.
41. Johnson EO, Hancock DB, Levy JL, Gaddis NC, Saccone NL, et al. (2013) Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy. Hum Genet 132: 509–522. doi:10.1007/s00439-013-1266-7.
42. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, et al. (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics 11: 587. doi:10.1186/1471-2105-11-587.
43. Siegmund KD (2011) Statistical approaches for the analysis of DNA methylation microarray data. Hum Genet 129: 585–595. doi:10.1007/s00439-011-0993-x.
44. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, et al. (2013) Review of processing and analysis methods for DNA methylation array data. Br J Cancer 109: 1394–1402. doi:10.1038/bjc.2013.496.
45. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, et al. (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics 13: 86. doi:10.1186/1471-2105-13-86.
46. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, et al. (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31: 142–147. doi:10.1038/nbt.2487.
47. Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15: R31. doi:10.1186/gb-2014-15-2-r31.
48. Koestler DC, Christensen B, Karagas MR, Marsit CJ, Langevin SM, et al. (2013) Blood-based profiles of DNA methylation predict the underlying

distribution of cell types: a validation analysis. Epigenetics Off J DNA Methylation Soc 8: 816–826. doi:10.4161/epi.25430.
49. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, et al. (2012) Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. PLoS ONE 7: e41361. doi:10.1371/journal.pone.0041361.
50. Mägi R, Morris AP (2010) GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics 11: 288. doi:10.1186/1471-2105-11-288.
51. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, et al. (2013) Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. Am J Hum Genet 93: 876–890. doi:10.1016/j.ajhg.2013.10.004.
52. Robert A, Eschwège V, Hameg H, Drouet L, Aillaud MF (1996) Anticoagulant response to Agkistrodon contortrix venom (ACV test): a new global test to screen for defects in the anticoagulant protein C pathway. Thromb Haemost 75: 562–566.
53. Dahlbäck B (1995) Factor V gene mutation causing inherited resistance to activated protein C as a basis for venous thromboembolism. J Intern Med 237: 221–227.
54. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, et al. (2011) Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet 43: 1091–1097. doi:10.1038/ng.946.
55. Weber M, Hellmann I, Stadler MB, Ramos L, Pääbo S, et al. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat Genet 39: 457–466. doi:10.1038/ng1990.
56. Gaidatzis D, Burger L, Murr R, Lerch A, Dessus-Babus S, et al. (2014) DNA sequence explains seemingly disordered methylation levels in partially methylated domains of Mammalian genomes. PLoS Genet 10: e1004143. doi:10.1371/journal.pgen.1004143.
57. Shoemaker R, Deng J, Wang W, Zhang K (2010) Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res 20: 883–889. doi:10.1101/gr.104695.109.
58. Kerkel K, Spadola A, Yuan E, Kosek J, Jiang L, et al. (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. Nat Genet 40: 904–908. doi:10.1038/ng.174.
59. Schlesinger F, Smith AD, Gingeras TR, Hannon GJ, Hodges E (2013) De novo DNA demethylation and noncoding transcription define active intergenic regulatory elements. Genome Res 23: 1601–1614. doi:10.1101/gr.157271.113.
60. Diaz GA, Banikazemi M, Oishi K, Desnick RJ, Gelb BD (1999) Mutations in a new gene encoding a thiamine transporter cause thiamine-responsive megaloblastic anaemia syndrome. Nat Genet 22: 309–312. doi:10.1038/10385.
61. Liu G, Yang F, Han B, Liu J, Nie G (2014) Identification of four SLC19A2 mutations in four Chinese thiamine responsive megaloblastic anemia patients without diabetes. Blood Cells Mol Dis 52: 203–204. doi:10.1016/j.bcmd.2013.11.002.
62. Wood MC, Tsiouris JA, Velinov M (2014) Recurrent psychiatric manifestations in thiamine-responsive megaloblastic anemia syndrome due to a novel mutation c.63_71 delACCGCTC in the gene SLC19A2. Psychiatry Clin Neurosci: n/a–n/a. doi:10.1111/pcn.12143.
63. Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392–404. doi:10.1038/nrg2579.
64. Thomas D (2010) Gene—environment-wide association studies: emerging approaches. Nat Rev Genet 11: 259–272. doi:10.1038/nrg2764.
65. Gagnon F, Aïssi D, Carrié A, Morange P-E, Trégouët D-A (2014) Robust validation of methylation levels association at CPT1A locus with lipid plasma levels. J Lipid Res. doi:10.1194/jlr.E051276.

5 Régulation épigénétique à longue distance par des variations génétiques

Dans la continuité du travail précédent qui illustrait que des polymorphismes génétiques d'un gène (*SLC19A2*) pouvaient influencer les niveaux de méthylation du gène dans lequel ils se situaient, j'ai également participé à une étude, en collaboration avec une équipe de l'institut de recherche en cancérologie d'Ontario (*Ontario Institute for Cancer Research*), visant à identifier des polymorphismes génétiques associés aux niveaux de méthylation de gènes plus éloignés de leur position génomique (Lemire *et al.* 2015). De tels polymorphismes sont communément appelés *trans* mSNPs (pour *methylation associated single nucleotide polymorphisms with trans effect*).

Pour cela, les niveaux de méthylation de 380 189 sites CpG ont été testés en relation avec 410 203 SNPs dans l'étude OFCCR (*Ontario Familial Colon Cancer Registry*) composée de 1 748 individus, pour moitié atteints de cancer colorectal. Contrairement aux études MARTHA et F5L-Pedigrees où les niveaux de méthylation ont été mesurés à partir de la totalité des cellules du sang périphérique, la méthylation dans l'étude OFCCR a été mesurée à partir des lymphocytes (ce qui représente environ 20 à 40% des leucocytes, cellules possédant de l'ADN dans le sang périphérique).

Parmi les 54 627 paires CpG-SNP identifiés dans l'étude OFCCR, 52 708 paires CpG-SNP concernaient des effets *cis*, c'est à dire que la distance entre le site CpG et le polymorphisme est inférieur à 1 Mb (mégabase). Les autres 1 919 paires CpG-SNP identifiées au seuil FDR 5% (ce qui correspond à une p-value de $3,2 * 10^{-13}$) concernaient des effets *trans*, dont 1 657 paires avec des sites CpG et polymorphismes localisés sur des chromosomes différents, et pour les 262 paires restantes sur le même chromosome mais à une distance supérieure à 1 Mb. J'ai eu en charge la réplication des 1 919 paires avec un effet *trans* dans l'étude MARTHA. 1 593 (83%) associations CpG-SNP furent répliquées dans l'étude MARTHA et 1 605 associations répliquèrent, au seuil FDR de 5%, dans MARTHA ou dans l'étude F5L-Pedigrees. Une étude d'enrichissement a ensuite été réalisée sur les mSNPs avec effet *trans*. Cette analyse a mis en évidence que ce type de polymorphismes était enrichi ($p < 0,001$) en polymorphismes situés dans des longs ARN non codant, dans des régulateurs épigénétiques connus et des réseaux d'ARN interagissant avec Piwi et autres facteurs de transcription dont les protéines à doigts de zinc.

Ces travaux montrent que des phénomènes complexes de régulations du niveau de méthylation de sites CpG par des polymorphismes se déroulent à travers tout le génome. Les paires SNP-CpG

identifiées dans ce travail peuvent servir à mieux comprendre les mécanismes d'interaction qui les caractérisent. La faible densité de polymorphismes (410 203) et de sites CpG (380 189) utilisés dans ce travail pour mettre en évidence ces phénomènes laisse envisager qu'il ne s'agit que de la partie émergée de l'iceberg.

nature COMMUNICATIONS

ARTICLE

Received 8 Oct 2014 | Accepted 19 Jan 2015 | Published 26 Feb 2015

DOI: 10.1038/ncomms7326 OPEN

Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci

Mathieu Lemire[1], Syed H.E. Zaidi[1], Maria Ban[2], Bing Ge[3], Dylan Aïssi[4,5,6], Marine Germain[4,5,6], Irfahan Kassam[7], Mike Wang[1], Brent W. Zanke[8], France Gagnon[7], Pierre-Emmanuel Morange[9,10,11], David-Alexandre Trégouët[4,5,6], Philip S. Wells[8], Stephen Sawcer[2], Steven Gallinger[12,13], Tomi Pastinen[3] & Thomas J. Hudson[1,14,15]

The interplay between genetic and epigenetic variation is only partially understood. One form of epigenetic variation is methylation at CpG sites, which can be measured as methylation quantitative trait loci (meQTL). Here we report that in a panel of lymphocytes from 1,748 individuals, methylation levels at 1,919 CpG sites are correlated with at least one distal (*trans*) single-nucleotide polymorphism (SNP) ($P<3.2\times10^{-13}$; FDR<5%). These *trans*-meQTLs include 1,657 SNP–CpG pairs from different chromosomes and 262 pairs from the same chromosome that are >1 Mb apart. Over 90% of these pairs are replicated (FDR<5%) in at least one of two independent data sets. Genomic loci harbouring *trans*-meQTLs are significantly enriched ($P<0.001$) for long non-coding transcripts (2.2-fold), known epigenetic regulators (2.3-fold), piwi-interacting RNA clusters (3.6-fold) and curated transcription factors (4.1-fold), including zinc-finger proteins (8.75-fold). Long-range epigenetic networks uncovered by this approach may be relevant to normal and disease states.

[1] Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 0A3. [2] Department of Clinical Neurosciences, University of Cambridge, Cambridge Biomedical Campus, Hills Road, Cambridge CB2 0QQ, UK. [3] McGill University and Genome Québec Innovation Centre, Montréal, Québec, Canada H3A 0G1. [4] INSERM, UMR-S 1166, Paris F-75013, France. [5] Sorbonne Universités, UPMC Univ Paris 06, UMR_S 1166, Team Genomics and Pathophysiology of Cardiovascular Diseases, Paris F-75013, France. [6] ICAN Institute for Cardiometabolism and Nutrition, Paris F-75013, France. [7] Dalla Lana School of Public Health, University of Toronto, College Street, Toronto, Ontario, Canada M5T 3M7. [8] Department of Medicine, University of Ottawa, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada K1H 8L6. [9] Laboratory of Haematology, La Timone Hospital, Marseille F-13385, France. [10] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Marseille F-13385, France. [11] Aix-Marseille University, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, Marseille F-13385, France. [12] Samuel Lunenfeld Research Institute, Toronto, Ontario, Canada M5S 1X5. [13] Division of General Surgery, Toronto General Hospital, Toronto, Ontario, Canada M5G 2C4. [14] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. [15] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada M5S 1A1. Correspondence and requests for materials should be addressed to M.L. (email: mathieu.lemire@oicr.on.ca).

Characterizing the relationships between genetic variants and functional elements of the genome is needed to advance our understanding of phenotypic diversity and characterize how genetic variation can perturb cell function and affect disease predisposition. Although cells contain a myriad of biomolecules that can potentially be assayed across large numbers of genetically characterized samples, analyses of nucleic elements such as transcript expression and CpG methylation are ideally suited for genome-wide comparisons with currently available technologies. In general, the detection of association between nearby (*cis*) genetic variants and functional elements has been easier to detect and validate than distal (*trans*) relationships, because of the multiple testing burden that arises when *trans*-associations are explored.

Expression quantitative trait loci (eQTL) studies in the past decade have demonstrated that gene transcript levels in a cell are frequently associated with nearby genetic variants (*cis*-eQTLs), which in turn are enriched for single-nucleotide polymorphisms (SNPs) associated with disease[1,2]. In parallel, mechanistic links between the disease-associated variants and the biology of the disease governed by local gene expression alterations are emerging[3]. The detection of eQTLs spanning long distances (*trans*-eQTLs) has been challenging, particularly due to the multiple testing burden that occurs when genome-wide sets of variants are compared with genome-wide sets of genes. To address this limitation, Westra *et al.*[4] performed an eQTL meta-analysis in 5,311 peripheral blood samples. They furthermore reduced the number of variants to test for *trans*-eQTLs to 4,542 SNPs that have been implicated in complex traits by genome-wide association studies. This allowed the detection and replication of *trans*-eQTLs for 233 disease-associated SNPs (at 103 independent loci), and provided insight into the pathogenesis of disease for selected variants identified in their screen.

An alternate approach to mapping regulatory relationships between genetic variants and distal genes is by correlating genetic variants with genome-wide epigenomic profiles. Notwithstanding rapid developments in epigenome-mapping methods that can explore a large number of chromatin modifications, the only approach that can screen thousands of samples needed to detect *trans*-acting relationships in a genome-wide fashion is an array-based multiplex assay that interrogates a multitude of methylation sites in parallel. A first-generation array that measured methylation levels at 22,290 CpG dinucleotides in 77 lymphoblastoid cell lines was used to detect associations between genome-wide SNPs and CpG methylation[5]. After applying a genome-wide false discovery rate (FDR) of 10% ($P = 2.1 \times 10^{-10}$), they detected 27 putative *cis*-association signals, most of which involved SNPs and CpG sites located within 50 kb of each other and 10 *trans*-association signals. Despite the limitations in the sample size used in this study, these preliminary studies demonstrated the potential to map *trans*-regulatory relationships between genetic variants and distal epigenetic elements that may affect gene regulation and complex disease phenotypes.

Here, we report a methylation quantitative trait loci (meQTL) study involving an analysis of methylation profiles for 380,189 CpGs and genotypes for 410,203 SNPs determined in lymphocytes from 898 patients with colon cancer and 850 controls, and subsequent replication in two independent data sets totalling 577 samples. To characterize the *trans*-meQTL loci that are statistically significant, we determined whether these loci are enriched in protein-coding and RNA-coding genes that could mediate the *trans*-acting effects. We explored *cis*- and *trans*-eQTLs correlating with meQTL loci in transcriptomic data sets derived from 137 CD4 and 137 CD8 lymphocytes. The analyses demonstrate an abundance of genetic loci that are associated with distal CpG methylation, a diversity of regulatory mechanisms that confer this role and networks of coordinated genes that are linked to biological and/or disease processes.

Results

Description of the data set. In this study, genome-wide SNP data and CpG methylation profiles detected using the Infinium HumanMethylation450 BeadChips were obtained from DNA extracted from lymphocytes of 898 patients with colon cancer and 850 controls (after sample exclusions, see Methods) enrolled in the Ontario Familial Colon Cancer Registry (OFCCR)[6,7]. We normalized the methylation data sets and identified and flagged sites with more than 1% missing data ($n = 13{,}032$), calculated after sample exclusions. We further identified CpG sites that are polymorphic irrespective of the minor allele frequency (MAF), sites for which at least one SNP with a MAF above 5% resides anywhere else underneath the probe sequence (69,974 sites derived from ref. 8) and sites that cross-hybridize with >90% identity to multiple regions (30,436 sites). These overlapping counts amount to 93,675 unique sites that were discarded. We retained the methylation values (also known as the β-values) at 380,189 autosomal CpG sites, as well as genotypes for 410,203 SNPs for further analyses. The OFCCR methylation and SNP data were deposited in dbGaP under the accession number phs000779.v1.p1.

For the data analyses that follow, we classify a SNP–CpG association as proximal if the SNP and the CpG site are separated by no more than 1 Mbp on the same homologous chromosome and as distal otherwise. Over 103 million SNP–CpG pairs were proximal candidates, and over 155 billion pairs were distal candidates. For each CpG site, we looked for SNPs that are associated with methylation levels in the 1,748 DNA samples; we note that we only combined cases and controls after performing separate analyses, which revealed over 80% overlap in significant results. The quantitative trait analyses were adjusted for sex, age, blood cell type surrogate (see Methods and Supplementary Fig. 1), batch (plate) and position of the sample on its array to get the significance and the proportion of the methylation variance explained by a SNP. We retained SNP–CpG associations consistent with an FDR (ref. 9) <5%, and report the effect size as the proportion R^2 of the CpG methylation variance that is explained by the SNP, among the variance not already explained by sex, age, blood cell type surrogate, batch and array position.

Proximal SNP–CpG associations. Owing to a rich body of literature[5,10–19], we present proximal SNP–CpG association results for completeness, but focus the rest of the study on distal (long range) associations.

Table 1 illustrates the distribution of the 52,708 CpG sites significantly associated with a proximal SNP (see also Supplementary Data 1), at an FDR <5%. At this FDR threshold, these CpG sites have at least 2.2% of their variance explained by at least one proximal SNP, with corresponding (unadjusted) significance level $P_{LRT} < 4.8 \times 10^{-10}$ derived from a likelihood ratio test (LRT). Following up these 52,708 sites by adjusting for additional covariates to account for hidden confounders (the top three principal components, that explain 92.5% of the variance) does not substantially alter this list: except for 313, all remain significant at FDR <5%; the 313 sites that failed this threshold nevertheless all have $P_{LRT} < 1.8 \times 10^{-6}$, with the large majority having $P_{LRT} < 10^{-8}$. The number of significant sites decreases with increasing variance explained; at 75% variance explained ($R^2 > 0.75$), the number of significant sites drops to 338. Table 1 further breaks down the set of CpG sites based on their variability, stratified in quintiles. As the methylation variability of the CpG site increases, the relative proportion of

sites correlating with a proximal SNP increases, as well as the proportion of the variance that they explain.

Distal SNP–CpG associations. We identified 1,919 CpG sites for which inter-individual methylation variations are significantly correlated with at least one distal SNP: 1,657 SNP–CpG pairs from non-homologous chromosomes and 262 SNP–CpG distant pairs (>1 Mb) located on the same homologous chromosome (Supplementary Data 2). These counts excluded CpG sites for which a proximal SNP was already identified. At an FDR <5%, these distal SNPs explain at least 3.1% of the variance of their companion CpG site, at a significance level $P_{LRT} < 3.2 \times 10^{-13}$. Following up these 1,919 sites by adjusting the analysis for additional covariates consisting of the top three principal components does not substantially alter this list: except for 33, all remain significant at FDR <5%; the 33 sites that failed to pass this threshold nevertheless all have $P_{LRT} < 1.5 \times 10^{-10}$. The case–control status is not a systematic or substantial confounder: at an FDR <5%, only 64 of the 1,919 sites (3.3%) display significant differences between cases and controls, no more than what is expected by chance alone. The very large majority (83.6%, or 1,605 pairs) of the 1,919 distal pairs are replicated in one (496 pairs) or both (1,109 pairs) replication sets (in MARTHA and/or F5L, both totalling 577 whole-blood samples) at FDR <5% (Supplementary Data 2), while replication of 8.5% of the pairs (165) was not attempted in both data sets due to poor quality of the intensity signals at the CpG sites or poor imputation quality ($R^2 < 0.3$) of the SNPs; ignoring the pairs that were not successfully attempted, the replication rate is thus 91.5%.

Similar to proximal pairs, where sites that showed greater inter-individual variability are more likely to be associated with a proximal SNP, the proportion of sites associated with a distal SNP also increases with the site's variability, but the increase is not as steep (Table 2).

The 1,387 unique SNPs (Supplementary Data 2) involved in the 1,919 significant SNP–CpG distal associations, as well as SNPs in linkage disequilibrium (LD) ($R^2 > 0.5$) with them, define the boundaries of a total of 1,074 non-overlapping genomic regions totalling 109.4 Mb. Figure 1 and Supplementary Fig. 2 illustrate the genomic landscape of meQTL loci. The regions surrounding the SNPs harbour 2,167 RefSeq genes (1,798 coding and 369 non-coding transcripts, including 314 long (>100 bp) non-coding transcripts). We observe enrichment for coding (1.7-fold), non-coding transcripts (2.0-fold) and long non-coding transcripts (2.2-fold). We compiled a list of 190 autosomal genes involved in epigenetic processes[20] (http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-085A.html; http://www.sabiosciences.com/rt_pcr_product/HTML/PAHS-086A.html) (Supplementary Data 3) and find 25 in our regions (13.2%, 2.3-fold enrichment). Of the 1,225 manually curated, high-confidence autosomal genomic loci that encode transcription factors (TFs)[21], 244 are found in the neighbourhood of these SNPs (19.9%, 4.1-fold enrichment). Of the 388 manually curated TF that are zinc-finger proteins (ZNFs), 155 are in our regions (39.9%, 8.75-fold enrichment). Also noteworthy in these regions are the presence of piRNA (piwi-interacting RNA; Genbank accession numbers (with DQ prefix) enumerated in ref. 22): 3,072 piRNA sequences are detected, typically in clusters, near these SNPs; this corresponds to a 3.6-fold enrichment compared with random regions of the genome. Repetitive elements such as short interspersed nucleotide elements (1.2-fold), and more specifically of the Alu family (1.5-fold), were also slightly enriched. Note that for all of the above features, the observed number of features per Mb exceeds the corresponding number of features in randomly selected regions in all 1,000 replicates, effectively providing point estimates of significance of $P < 0.001$

Table 1 | Number of CpG sites significantly associated with at least one proximal SNP.

R^2	All sites	Q1 (0–2.1%)	Q2 (2.1–4.7%)	Q3 (4.7–10.9%)	Q4 (10.9–21.4%)	Q5 (21.4–100%)
2–5%	19,216	769	1,775	3,521	5,920	7,229
5–10%	14,494	355	951	2,234	4,508	6,444
10–25%	12,284	164	515	1,498	3,490	6,616
25–50%	4,960	22	89	370	1,133	3,345
50–75%	1,416	2	5	30	225	1,154
75%+	338	0	0	1	8	329
Total	52,708	1,312	3,335	7,654	15,284	25,117

SNP, single-nucleotide polymorphism.
Counts are stratified based on the site's inter-individual variability and R^2-value (proportion of methylation variance explained) between the SNP and CpG pairs. The variability of a site is measured as its 95%-reference range (the difference between the most and least methylated individuals, among 95% of the individuals forming the central distribution of methylation values) stratified in quintiles (Q1–Q5); percentages in brackets indicate the corresponding 95%-reference range values.

Table 2 | Number of CpG sites significantly associated with at least one distal SNP.

R^2	All sites	Q1 (0–4.9%)	Q2 (4.9–7.0%)	Q3 (7.0–10.8%)	Q4 (10.8–17.8%)	Q5 (17.8–100%)
2–5%	567	80	104	106	123	154
5–10%	676	68	122	154	149	183
10–25%	441	29	66	111	116	119
25–50%	177	6	38	38	45	50
50–75%	45	0	2	6	13	24
75%+	13	0	0	1	3	9
Total	1,919	183	332	416	449	539

SNP, single-nucleotide polymorphism.
Counts are stratified based on the site's inter-individual variability and R^2-value (proportion of methylation variance explained) between the SNP and CpG pairs. The variability of a site is measured as its 95%-reference range (the difference between the most and least methylated individuals, among 95% of the individuals forming the central distribution of methylation values) stratified in quintiles (Q1–Q5); percentages in brackets indicate the corresponding 95%-reference range values.

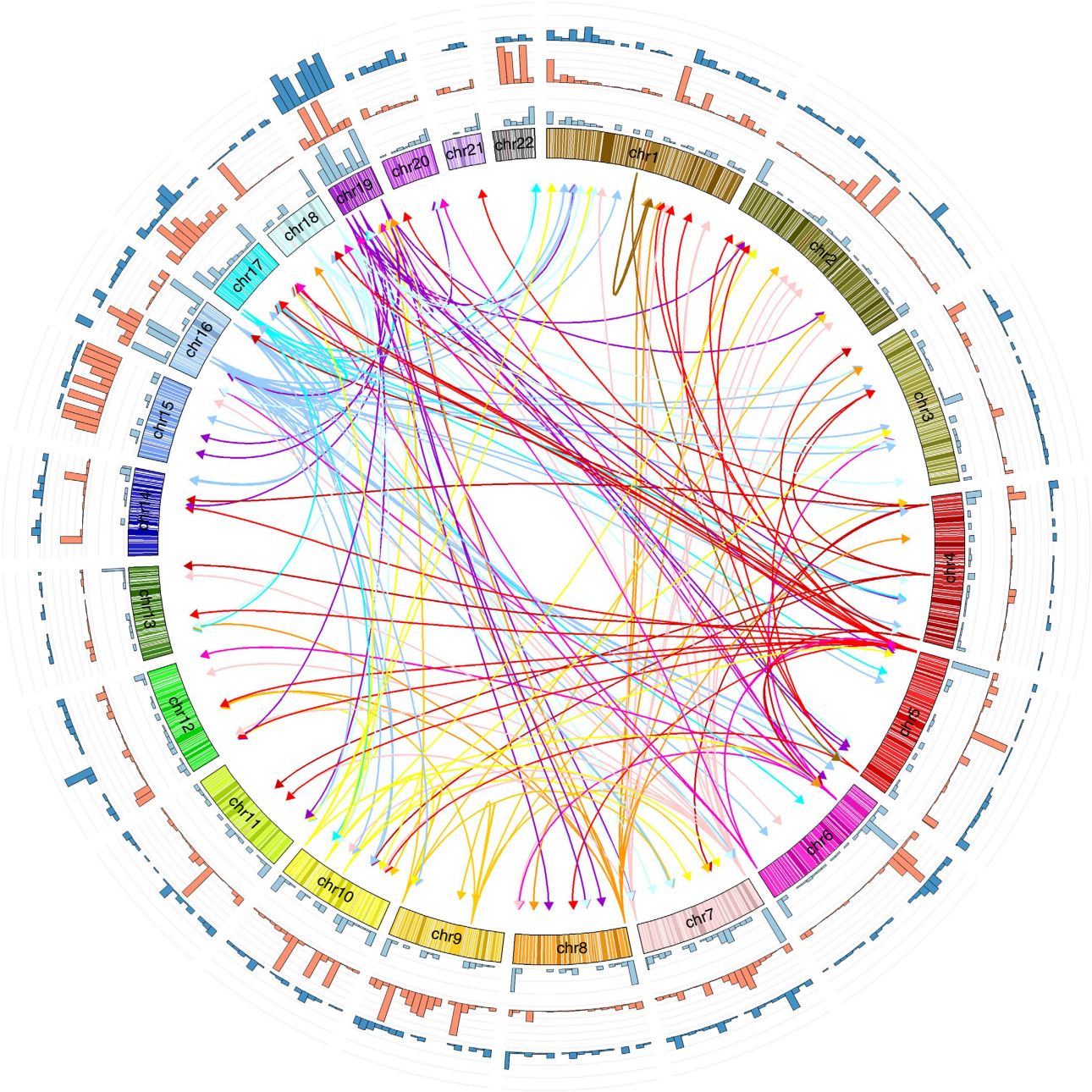

Figure 1 | Enrichment of features in regions harbouring SNPs involved in distal SNP–CpG associations. Outer histograms: number of SNPs involved in distal SNP–CpG associations (light blue), calculated in 7.5 Mb bins; number of piRNA sequences (orange); number of transcription factors (dark blue). Inner links: SNP regions associated with four or more CpG sites. Arrows are pointing from SNPs to the CpG sites they are associated with, and are coloured according to the chromosomes where the SNPs reside.

(estimated from proportions) for each of the above absolute fold enrichments. We also observe that 195 (18.1%) of the 1,074 non-overlapping genomic regions harbouring genetic variants associated with meQTLs are located towards the subtelomeric regions of the chromosomes (within 2 Mbp of the chromosome's ends), as well as several small genomic regions containing one or more SNPs associated with multiple CpG sites.

Genes in the SNP LD-defined neighbourhoods and genes annotated to the CpG sites they are associated with were projected into functional interaction (FI) networks[23] (a high-throughput network of protein–protein interactions) to evaluate which pairs display protein–protein interactions, or to see whether genes involved in distal SNP–CpG pairs are closer in the FI network than expected by chance. Out of the 4,131 gene–gene pairs that can be formed by joining one SNP-annotated gene and one of its associated CpG-annotated genes, 1,140 pairs had both genes mapped to FI networks. The shortest path between these pairs of genes was calculated to be 3.48 (meaning that they are connected but only through another 2.48 genes on average), no different from what is expected from random pairs of connected genes ($\sim$3.50, estimated from 1,000 replicates, $P = 0.2$ estimated from proportions). Only two pairs of genes display direct protein–protein interactions: UBE2N (annotated to rs6538421) and UBR4 (annotated to cg00223950); and ERBB4 (annotated to rs7599312) and SH3GLB2 (annotated to cg20548744).

Expression quantitative loci associated with distal meQTLs. We considered the possibility that distal sites with differential CpG methylation could display differential allelic expression of transcripts that are in the vicinity of the CpG. We analysed independent genetic and transcriptomic data sets derived from 137 CD4 and 137 CD8 lymphocytes for the presence of eQTLs associated with the 1,387 unique SNPs involved in distal SNP–CpG methylations.

Figure 2a,b displays quantile–quantile plots of significance levels against theoretical quantiles for associations between the SNPs and expression levels of distal protein-coding and non-coding transcripts found in their LD-defined neighbourhoods, in CD4 and CD8 cells. In each cell type, the weight of the distribution deviates towards greater levels of associations than predicted by chance, supporting the presence of enrichment of *trans*-eQTLs. Supplementary Data 2 lists the results of all tested *trans*-eQTL/gene pairs, and also present *cis* pairs for completeness.

In Battle *et al.*[24], *trans*-eQTL (interchromosomal) were identified from a larger number of samples (922 DNA samples derived from whole blood) for 138 genes and distant eQTL (intrachromosomal eQTL, >1 Mb from the gene) for 269 genes, with 5 genes overlapping both lists. Six of these genes are also displaying significant meQTL with at least one annotated CpG site, either with the SNP itself or one in LD with it (smallest observed $R^2>0.70$). These are: *PDE4DIP* (annotated to a distant eQTL), *DUSP22*, *PGLS*, *ZNF154*, *ZNF274* and *ZNF551* (annotated to *trans*-eQTLs).

Relevance of meQTLs with respect to autoimmune diseases. We have compared the list of proximal and distal meQTLs with autoimmune disease (AID)-associated SNPs. Using Immunobase (https://www.immunobase.org), we generated a list of the currently known associated SNPs (non-major histocompatibility complex). In total, we found 552 associations across 11 AID (indexed in Immunobase; considering Crohn's disease and ulcerative colitis together as inflammatory bowel diseases) involving 512 unique SNPs. Furthermore, since some of the associated SNPs in one disease are in high LD ($R^2>0.8$) with associated SNPs in other diseases, this list covers a total of 447 associated loci (two are on the X chromosome). These 512 AID-associated SNPs overlap or are highly correlated ($R^2>0.8$) with 200 of the proximal meQTLs (associated with a total of 561 CpG sites) and 14 of the distal meQTLs (associated with a total of 24 CpG sites) (Supplementary Data 4). Alternatively, 161/552 (29%) of the reported AID associations overlap or correlate with at least one proximal meQTL, and 9/552 (2%) overlap or correlate with at least one of the distal SNPs. In terms of the 447 independent loci, these percentages are, respectively, 115/447 (26%) and 8/447 (2%). This descriptive analysis illustrates that localization of disease-associated genes can benefit from the meQTL annotation of disease-associated SNPs.

Case studies of three *trans*-meQTL loci. SENP7 (Sentrin/small ubiquitin-like modifier (SUMO)-specific protease 7) is an isopeptidase, which catalyses the deSUMOylation of SUMO2/3-conjugated proteins. The reversible post-translational SUMO modifications of proteins regulate many cellular processes including replication, transcription, recombination, chromosome segregation and cytokinesis. SENP7 interacts with BCL6, CBX5, KAP1 and HP1 alpha proteins, which are involved in epigenetic repression. SENP7 was recently described to promote chromatin relaxation in response to DNA damage, for repair and for cellular resistance to DNA-damaging agents[25]. Depletion of SENP7 results in spread of heterochromatin factors and condensed chromatin[25].

In our data set, four intronic SNPs (rs2553419, rs2682386, rs9859077 and rs2141180) of the *SENP7* gene in high LD with each other correlate with the methylation levels at numerous CpG sites (replicated in at least MARTHA or F5L for the large majority; Supplementary Data 2); these SNPs correlate with *cis*-acting regulation of *SENP7* expression in CD4 and CD8 lymphocytes (FDR<5%; $P_{Wald}<0.00015$ derived from a Wald test; Supplementary Data 2) and *trans*-acting regulation of several distal genes, including *ZNF154*, *ZNF274* and *ZNF814* (Fig. 3; Supplementary Data 2), which reside within a ~250-kb region on chromosome 19, as well as *ZNF268* (chromosome 12) and *LDHD* (chromosome 16). Elevated levels of *SENP7* transcripts are associated with reduced CpG methylation located in 5′-flanking regions of *ZNF154*, *ZNF274* and *ZNF814* and increased transcript levels of these genes ($P_{Wald}<0.006$); other genes whose CpG methylation is associated with *SENP7* intronic SNPs show similar trends of expression changes that do not meet the significance threshold (FDR>5%). These data indicate that SENP7 transcript levels regulate methylation and transcript expression of *trans*-meQTLs at multiple sites.

The identification of *SENP7*-regulated genes such as *ZNF154* potentially extends the understanding of SENP7 function and linkages to cancer processes. ZNF154, a putative TF expressed in many tissues, was initially identified as commonly deleted in thyroid adenoma[26]. Hypermethylation at the CpG island in the promoter region of *ZNF154*, and negative correlation between methylation and gene expression were recently described in serous ovarian cancers, and in endometrioid ovarian and endometrial cancers[27]. Methylation-mediated repression of

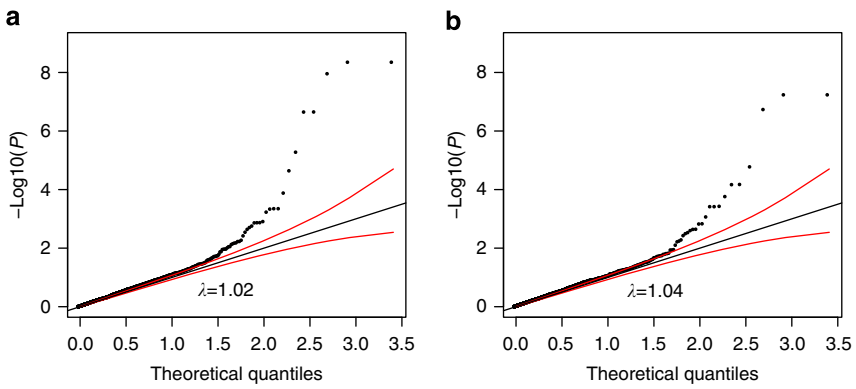

Figure 2 | Quantile–quantile plot of association levels of *trans*-eQTL. (**a**) *Trans*-eQTL in CD4; (**b**) *trans*-eQTL in CD8. Red lines are 95% confidence bands. λ is the inflation factor, the ratio of observed to expected medians.

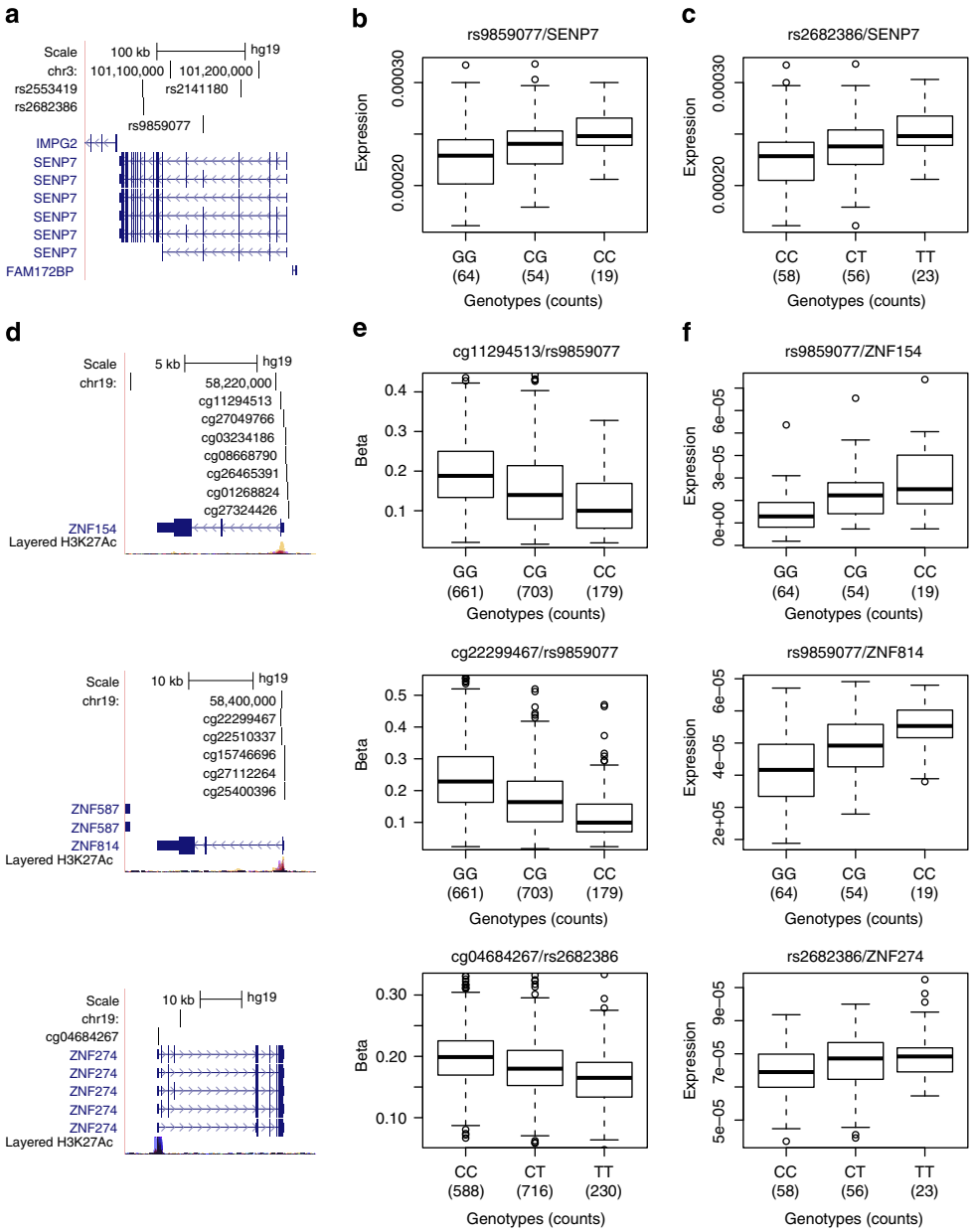

Figure 3 | The *SENP7* locus and its distally associated CpG sites. (**a**) UCSC browser illustration of the gene structure and location of SNPs associated with distal CpG sites. (**b**,**c**) Association in *cis* between expression of *SENP7* in CD4 cells and (**b**) rs9859077, (**c**) rs2682386. (**d**) UCSC browser illustrations of selected associated distal CpG sites and their annotated genes (from top to bottom: *ZNF154, ZNF814* and *ZNF274*). (**e**) Corresponding boxplot representations of the SNP/CpG associations. (**f**) Corresponding boxplot representations of the association in *trans* between expression of the genes in CD4 cells with SNPs.

ZNF154 in ovarian cancer is associated with poor overall survival[28]. In clear cell renal cell carcinoma[29] and in bladder cancer *ZNF154* is hypermethylated, and *ZNF154* methylation is identified as a biomarker of bladder cancer recurrence[30,31].

A chromosome 16 region of LD with rs12933229 ($R^2 > 0.5$), which is devoid of protein-coding genes, is rich in piRNAs and contains *RRN3P2* (a long non-coding RNA gene), is a *trans*-meQTL locus associated with methylation at 20 distal CpG sites (all replicated in both MARTHA and F5L) residing on nine non-homologous chromosomes (Fig. 4 for selected genes whose expression are the most strongly associated with the SNP; Supplementary Data 5). Five of these *trans*-meQTL-associated regions have independently been reported[19]. Enrichment of the H3K27Ac histone marks and TF-binding sites as determined by Chip-seq assays[32], as well as the presence of spliced expressed sequence tags (ESTs) indicate enhanced transcriptional activity in the region encompassing the piRNA cluster (Supplementary Fig. 3). piRNA clusters are usually transcribed as long single-stranded RNA and processed into mature piRNAs of 25–33 nucleotides in length[33,34]. The rs12933229 also correlates with *cis*-acting regulation of *RRN3P2* expression in CD4 and CD8 lymphocytes ($P_{Wald} = 0.0012$ and $P_{Wald} = 0.038$, respectively). Long non-coding RNA (linc) *cis*-eQTLs can also influence the expression levels of downstream genes[35]; in this case, *RRN3P2* could be a *cis*-regulator of piRNA transcription. Due to the described role of piRNA in methylation, we surmise that the *trans*-regulatory effects on methylation at distal genes may be mediated by one or more of the 21 piRNA genes (which were not

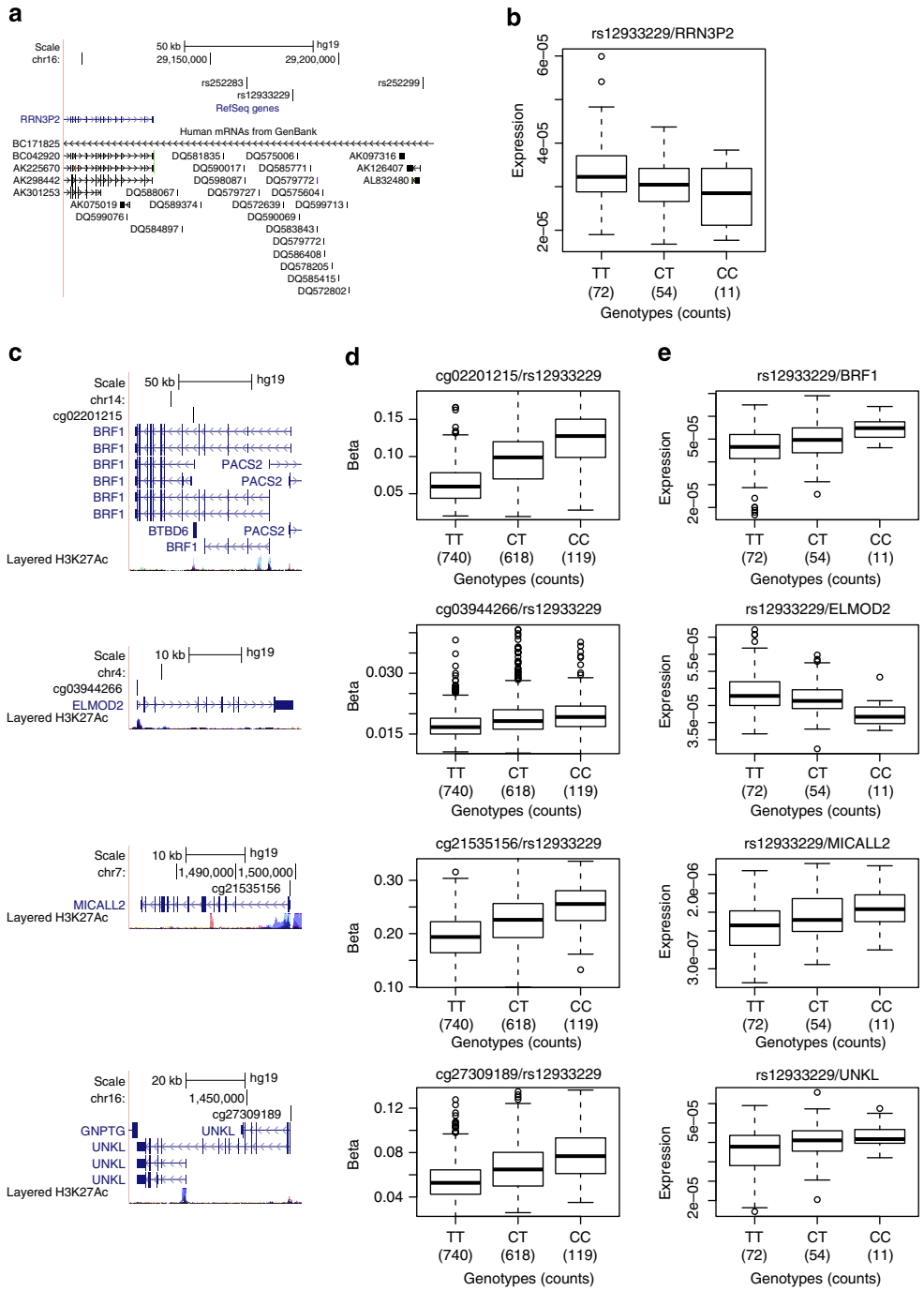

Figure 4 | The *RRN3P2* locus and its distally associated CpG sites. (**a**) UCSC browser illustration of the gene structure and location of SNPs associated with distal CpG sites. (**b**) Association in *cis* between expression of *RRN3P2* in CD4 cells and rs12933229. (**c**) UCSC browser illustrations of selected associated distal CpG sites and their annotated genes (from top to bottom: *BRF1*, *ELMOD2*, *MICALL2* and *UNKL*). (**d**) Corresponding boxplot representations of the SNP/CpG associations. (**e**) Corresponding boxplot representations of the association in *trans* between expression of the genes in CD4 cells with rs12933229.

measured in the lymphocyte RNA expression studies) in the ~56-kb region harbouring SNPs in LD with rs12933229. The Piwi/piRNA pathway has been described to promote heterochromatin formation, DNA methylation, transcriptional and post-transcriptional gene silencing and transcriptional activation by promoting euchromatic features[34,36,37].

We observe that for 19/20 (95%) of distal CpG sites that are associated with rs12933229, the β-value increases with each copy of the C allele at rs12933229, suggesting that the *trans*-effect is consistent in how it distally regulates methylation (Supplementary Data 5). Interestingly, there are nine transcripts in the vicinity of these distal CpGs that show expression level changes in CD4 cells that are associated with rs12933229 ($P_{Wald}<5\%$) (Supplementary Data 5). Two genes (*ELMOD2* and *SLC35A3*) show increased expression correlating with hypomethylation of their respective CpGs (cg03944266 and cg16383001). However, we observe increased gene expression correlating with CpG hypermethylation for seven genes (*STARD3*/cg01325958, *BRF1*/cg02201215,

BTBD6/cg02201215, *MAP3K14*/cg07370464, *GPC2*/cg13921324, *MICALL2*/cg21535156 and *UNKL*/cg27309189). This unusual association of increased CpG methylation and increased expression of local transcripts does not follow the classical 'negative correlations' between DNA methylation and gene expression. Recent studies comparing DNA methylation and gene expression have reported this paradox[5,14,18,38], and surmised that the position of the CpG within the gene body may be relevant, with DNA methylation in promoters being negatively correlated, while DNA methylation within the gene body[39] or 3′untranslated regions[40] being positively correlated with gene expression, through mechanisms not yet explained. In our data set, we do not observe consistent patterns between the position of the CpG and the relationships between CpG methylation and gene expression. Further efforts are needed to tease out these effects resulting from a number of possible mechanisms involving chromatin state, histone modifications, TF availability and binding, as well as post-transcriptional regulation, including changes in RNA decay and stability by RNA-binding proteins and levels of small RNAs.

CTCF (CCCTC-binding factor) is a zinc-finger DNA-binding protein that functions in transcription (activation and repression), RNA splicing, insulator activity, chromatin architecture and in genomic imprinting[41]. It is estimated that there are ~30,000 CTCF-binding sites in the human genome[42]. The rs7203742 SNP in the intron of *CTCF* is associated with methylation levels at 14 CpG sites across the genome (all replicated in MARTHA, the majority replicated in F5L; Supplementary Data 2). For *AOC2/PSME3*, *SEC14L1*, *PGLYRP2*, *ETS1*, *C1S*, *GPSM1* and *SLC26A11* genes (Fig. 5), and in regions without an annotated transcript, the rs7203742-associated CpG sites overlaps with CTCF-binding sites. This is consistent with observations that TFs can influence methylation at their binding sites[18,43]. These CpG sites reside in the promoter regions, introns and exons (untranslated and coding) of the associated genes. The CTCF binding to non-methylated/hypomethylated CpG sites, and its inhibition of binding to methylated DNA in concert with TFs and distal enhancers could result in repression or activation of the resident transcript. Such CTCF-mediated transcriptional repression and activation have been described for the imprinting control region of *Igf2/H19* genes[41,44].

In both the CD4 and CD8 lymphocyte expression data sets described above, we did not detect significant eQTLs associations.

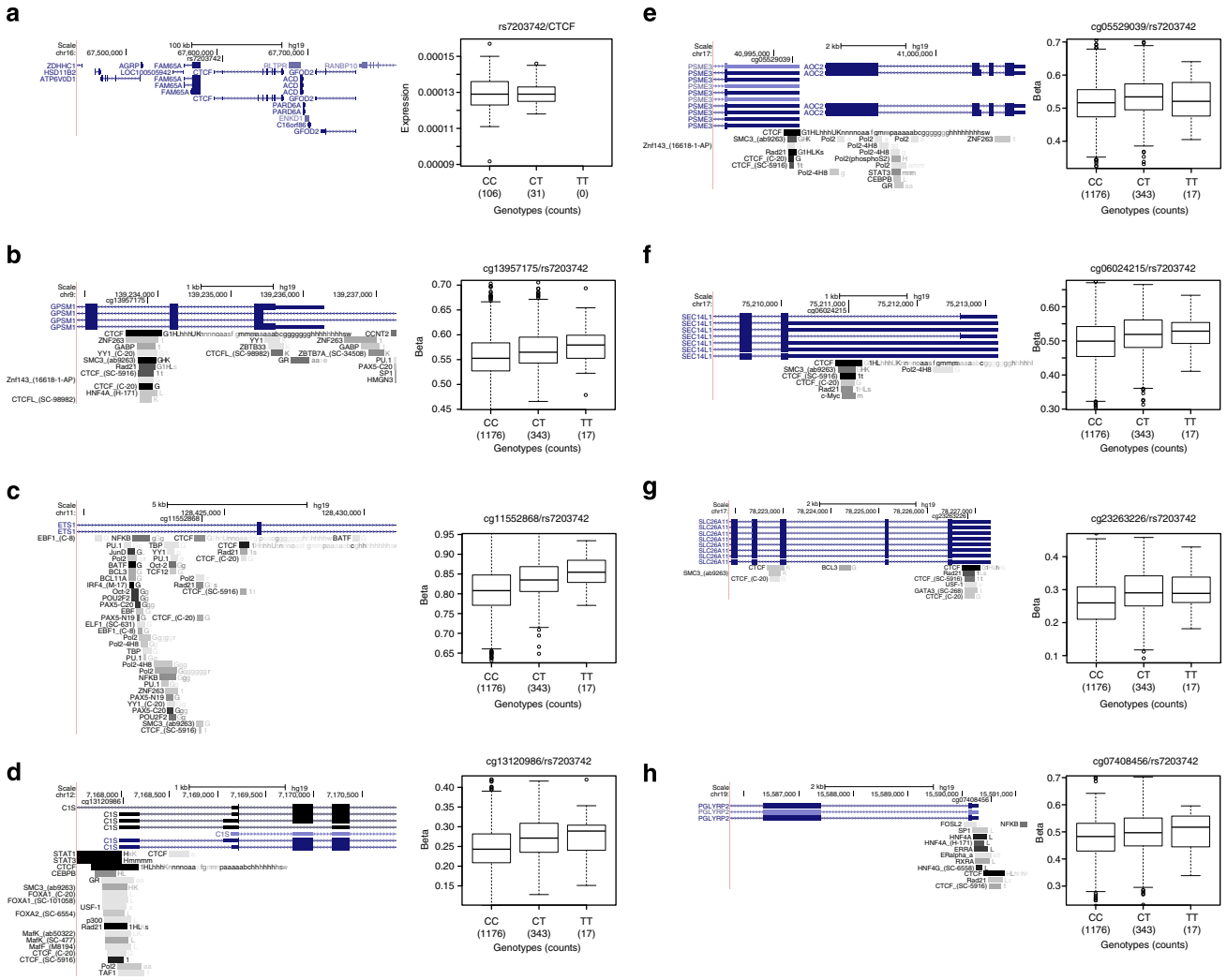

Figure 5 | The *CTCF* locus and its distally associated CpG sites. (**a**) UCSC browser illustration of the gene structure and location of SNPs associated with distal CpG sites, and boxplot representation of the association in *cis* between CTCF expression and SNP genotypes. (**b**–**h**) Distally associated CpG sites, their gene annotations and boxplot representations of SNP–CpG associations. The Transcription Factor ChIP-seq from ENCODE tracks are displayed in the UCSC browser illustrations to indicate CTCF-binding sites.

However, there are published reports that independently validate CTCF-related meQTLs. Transcriptome sequencing on lymphocyte messenger RNA from patients with intellectual disability (MIM 604167), harbouring mutant CTCF, and healthy individuals identified *SEC14L1* as the top-ranked dysregulated gene in patients[45]. Conditional deletion of *Ctcf* in mice myeloid cells resulted in 4.77-fold upregulation of *C1s* gene in *Ctcf*-deficient macrophages[46]. These data substantiate the repression and activation effects of CTCF on the expression of *SEC14L1* and *C1s* genes, respectively.

Discussion

In this study, we demonstrate that there is extensive long-range regulation of CpG methylation associated with genetic variation in the genome. A strength of the study is the large sample size used for the discovery of meQTL loci. The vast majority of our distal associations were replicated independently, using a variety of sampling designs (case–controls, case-only and family based), disease (colon cancer or venous thrombosis (VT)) normalization and statistical framework (linear models with fixed-only or mixed effects; using the methylation values as the dependent variable or their logit-tranformations; using LRTs or robust Wald tests) and cell type heterogeneity correction (principal component analysis; measured); they are statistically robust. While this data set is neither comprehensive due the incompleteness of the CpG sites that were ascertained and the single source of tissue that was used, the implications of the results are daunting, as it is quite likely that the $>1{,}000$ meQTLs represent the tip of the iceberg in regards to the numbers of distal pairs linking genetic variants and with CpG methylation that are present in the human genome. Revisiting the data sets presented here using more powerful strategies or combining them, as well as other data sets, by means of meta-analyses would likely reveal other meQTLs and their distal targets.

We were able to further analyse our meQTL results in expression data sets derived in an independent set of genotyped lymphocytes. This allowed us to identify many genetic variants that are associated with both differential CpG methylation and expression affecting distal genes. We note that future studies would benefit from having methylation and expression data sets generated in the same tissue panels, as it could enhance the detection of interrelated effects on transcription and expression. Nonetheless, the current data sets provided numerous examples of this, with *SENP7* being a prime example.

We focused our analyses on loci that correlate with CpG methylation at multiple distal sites. This revealed a number of possible regulators including SENP7, CTCF and a piRNA cluster, although the putative mechanism of *trans*-regulation appears to be different for each. While the current concepts on how each of the regulators may affect distal regulation is only partially known, the widespread functions that these genes are reported to have on gene transcription and chromatin architecture are consistent with our observations.

Finally, this study provides new avenues to understand molecular processes that are coordinated by genetic and epigenetic mechanisms, and study genetic loci associated with disease predisposition.

Methods

Methylation sample description. The cases and controls used in this study were enrolled in phase I of the OFCCR[6,47]. Briefly, probands were selected from incident colorectal cancer cases identified between 1 July 1997 and 30 June 2000 from the population-based Ontario Cancer Registry. Controls were recruited through telephone interviews in randomly selected households in Ontario. A subset of OFCCR cases and controls were used in the Assessment of Risk for Colorectal Cancer Tumours in Canada study that led to the identification of a genetic risk factor for colon cancer in the 8q24 region[7] and are used in this study. Informed consent was obtained from all participants.

Methylation profiling. We profiled 2,203 samples from 2,101 unique donors (1,103 cases of colorectal cancer and 998 controls). Lymphocyte pellets were extracted from whole blood using Ficoll-Paque PLUS (GE Healthcare). DNA was extracted from lymphocytes using phenol–chloroform or the Qiagen Mini-Amp DNA kit, except for 99 samples (90 cases, 9 controls), for which DNA was extracted from lymphoblastoid cell lines. We used 15 µl of DNA from all samples at concentrations averaging 90 ng µl^{-1} (20 ng µl^{-1} s.e.). DNA samples were bisulfite-converted using the EZ-96 DNA Methylation-Gold Kit (Zymo Research, Orange, CA); 4 µl of bisulfite-treated DNA was then analysed on the HumanMethylation450 BeadChip from Illumina according to the manufacturer's protocol. Except for the case/control status, the plating of the DNA samples was not based on any specific characteristics that they might share (for example, gender and age), effectively randomizing these factors on the arrays. Most cases and most controls were plated and processed separately, except for 125 cases that were matched to 125 controls on 125 rows of 36 arrays; for these samples, batch and position effects are controlled for by design.

Calculation of methylation ratios from intensities. We calculated the methylation ratios (proportion of molecules that are methylated at a given site, also known as β-value) as $\beta = \min(M,1)/(\min(M,1) + \min(U,1))$ where M and U are, respectively, the methylated and unmethylated intensity signals (possibly background-corrected signals). This is a slight modification of GenomeStudio's calculation: Illumina defines the ratio as $\beta = \min(M,0)/(\min(M,0) + \min(U,0) + 100)$. In this expression, the offset (100) artificially moves the methylation value away from 1. Moreover, the possibility of the numerator being 0 makes the resulting distribution of β-values both discrete and continuous (with point mass at 0). We prefer our expression for β, which results in a continuous distribution that does not bias against a fully methylated state.

Processing of idat files and normalization. We used functions from the methylumi package from Bioconductor to read the idat files and write intensity values to text files. We compared the methylation values (the proportion of molecules that are methylated at a given site, also known as β-value) of a number of normalization strategies by focusing on 64 samples that were done in duplicate or triplicate (129 possible pairs after sample exclusions) (see Supplementary Methods and Supplementary Figs 4–11). On the basis of these comparisons, we hereafter use methylation values derived from data that was background-corrected using NOOB[48] followed by colour adjustments using Illumina's normalization probes and algorithms; BMIQ[49] was then applied to the set of β-values.

Principal component analysis. We performed a principal component analysis by first calculating the covariance matrix between all samples using only the most variable autosomal CpG sites, measured in terms of their 95% reference range: the range of methylation values observed in the central 95% of the samples, or more precisely the difference between the 97.5 and 2.5% percentiles. Using a 95% reference range of at least 0.20, 131,045 CpG sites were used in the covariance matrix calculation. Together, the top three principal components explain over 92.5% of the total variance. Each subsequent vector does not add substantially to the variance explained: 98 vectors would be necessary to explain 95% of the total variance.

Sample exclusion. We excluded from association analyses: (1) samples for which DNA was extracted from lymphoblastoid cell lines (99 samples including 1 in duplicate); (2) samples that were outliers with respect to any one of the internal control probes (excluding probes designed to evaluate the background noise and probes designed to normalize the data), where an outlier is defined as any value more than three times the interquartile range away from the closest quartile (124 samples); (3) samples that were outliers with respect to any one of the first 12 principal components (corresponding to the approximate location of the elbow of the eigenvalue scree plot), recomputed after exclusion of the samples in steps 1 and 2 (26 samples; see Supplementary Methods); (4) samples that were not of non-Hispanic white ancestry, either self-declared or by investigation of genetic ancestry using genome-wide SNP data (164 samples); and (5) samples with intensities on the X or Y chromosome inconsistent with their gender (seven samples). These overlapping counts sum up to 389 unique samples that were excluded. For duplicate samples, we only kept the one with the smallest number of sites with a detection P value (see below) less than 0.01. After exclusion of samples, we were left with 1,748 samples: 898 cases (females: 58.9%; mean age: 63 ± 8.1 s.d.) and 850 controls (females: 42%; mean age: 64.3 ± 8.2 s.d.).

Probe exclusion. The detection P value is a quantitative assessment of the probability that the total intensity, for a given sample at a given site, can be distinguished from background noise, and is included along with the intensity values. All β-values with a detection P value less than 1% were treated as missing data. Sites with more than 1% missing values after sample exclusion were discarded (13,032 sites).

Polymorphic sites. We excluded from SNP–CpG-methylation association analyses all CpG sites that were polymorphic at the cytosine or at the guanine base, and, in the case of Infinium I probes, had a SNP at the position where single-base extension occurs (the base before the CpG cytosine), irrespective of the allele frequency. We moreover excluded CpG sites for which a SNP resided elsewhere within the probe target sequence as long as its MAF was above 5%. Allele frequency estimates were extracted from the samples of European ancestry (EUR; 379 samples) of the 1000 Genomes Project[50].

Probe cross-reactivity. We excluded CpG sites from SNP–CpG-methylation association analyses if their probe sequences aligned to multiple positions with $\geq$90% identity; misalignments were ignored if the terminal nucleotide of the probe was a mismatch (preventing single-base extension), if the probe aligned with gaps or if the probe mapped onto an alternative assembly of the target chromosome. See ref. 8 for additional details.

Cell-type proportions. Because differences in cell-type proportions between DNA samples can confound association results[51], we adjusted our analyses using a surrogate for cell-type proportions derived from 49 differentially methylated CpG sites present on the HumanMehtylation450 array that have the ability of discriminating between blood cell types[52]. As a surrogate for cell-type proportions, and to reduce the number of variables, we used the first two principal components associated with these 49 sites that together explain over 90% of the total variance.

To verify that the first two principal components that we derived from the list of 49 differentially methylated CpG sites[52] can indeed serve as a surrogate for blood cell proportions, we tested for associations between the principal components and the methylation levels at all of our sites, adjusting our analyses for sex, age, arrays and position of the samples on the arrays. We selected the top 10% of the sites that showed the strongest associations (46,000 sites, all associated at levels $P<10^{-100}$) and extracted these sites in data sets of purified human leukocyte subtypes[53] (GEO accession: GSE39981); 2,605 sites were overlapping. A dendrogram representation of our top sites in this data set[53] reveals clear clustering of samples according to cell types, indicating a good ability for principal components to discriminate between samples with different cell compositions (Supplementary Fig. 1).

Genetic variants and inter-individual levels of methylation. For each CpG site, we looked for SNPs that were associated with inter-individual methylation levels; we treated the methylation levels as a quantitative trait, and searched the genome for quantitative trait loci (meQTL) that explained a substantial proportion of the methylation variance at that site. We used SNP data from a colon cancer genome-wide association study[7] (a 1536 GoldenGate panel from Illumina; the 10-k coding-SNP array from Affymetrix/ParAllele; the Human Mapping 100-k set from Affymetrix) all combined and complemented with the Human Mapping 500-k set from Affymetrix. When combining SNPs genotyped on multiple platforms or arrays, SNPs with more than 0.5% discordant calls were discarded; otherwise discordant calls were considered to be missing calls. SNPs with minor allele frequencies less than 5% in either the cases or the controls were ignored, as well as SNPs with genotypic frequencies inconsistent with Hardy–Weinberg equilibrium at $P<10^{-4}$ in all samples and SNPs with call rates less than 90%; in all, 410,203 autosomal SNPs were available for pairwise comparison with 380,189 autosomal methylation sites.

We classify a SNP–CpG association as proximal if the SNP and the CpG site are separated by no more than 1 Mbp on the same homologous chromosome and as distal otherwise. Over 103 million SNP–CpG pairs were proximal candidates, while over 155 billion pairs were distal candidates. Preliminary analyses in cases and controls separately revealed over 80% overlap in significant results at stringent significance thresholds; we thus combined cases and controls to increase power.

We analysed the SNP–CpG pairs using a linear regression framework implemented in R: the β-value of a methylation site was taken as the independent variable, genotypes (number of minor alleles) at a SNP were taken as the values of the independent variable and the model included sex, age, case/control status, cell-type proportion surrogate, batch (plate) and position of the sample on its array as covariates. Significance of the SNP was calculated from a LRT (P_{LRT}). Rare homozygous genotypes (count of less than 10) were combined with heterozygotes. Following this analysis, we report SNP–CpG associations consistent with an FDR (ref. 9) $<$5%, and report the effect size as the proportion R^2 of the CpG methylation variance that is explained by the SNP, among the variance not already explained by the covariates.

We stratify results based on the inter-individual variability of the CpG sites. As a measure of variability, we used the 95%-reference range (the difference between the most and least methylated individuals, among 95% of the individual forming the central distribution of methylation values), which is less sensitive to outliers than the full range and more readily interpretable than the s.d.

All genomic positions are reported with respect to the hg19/GRCh37 coordinates.

A circular representation of distal SNP–CpG associations with ideograms was plotted using CIRCOS[54].

SNP–CpG replication sets. We sought to replicate distal SNP–CpG associations using two sets. (1) F5L: a set of 227 French–Canadian individuals belonging to five extended families (each ascertained through a single proband with idiopathic VT and carrying the factor V Leiden mutation[55]. These whole-blood samples were profiled on the HumanMethylation450 array and genotyped on the Human660W-Quad array from Illumina. We chose proxy SNPs from the Human660W-Quad array that are in high LD with SNPs involved in distal SNP–CpG associations when the latter were not present on the Illumina array; to seek consistency in the direction of association, we report phasing information and strength of LD between the pairs of SNPs (based on 1000 Genomes EUR data). (2) MARTHA: a set of 350 unrelated cases of VT patients of European ancestry (primarily of French descent) from the MARseille THrombosis Association study[55,56]. These whole-blood samples were profiled on the HumanMethylation450 array and genotyped on the Human610-Quad array from Illumina. Imputation of ungenotyped SNPs was performed using MACH v1.0.16a and minimac v4.4.3 (ref. 57) with their accompanying 1000 Genomes-based reference set[58]. We report the R^2 measure of imputation accuracy.

All three sets (OFCCR, F5L and MARTHA) were analysed independently by the different groups providing the data, using methods and software seen as most appropriate for their own samples. Both F5L and MARTHA methylation data were normalized similarly to the OFCCR data set, with the exception that SWAN[59] was used instead of BMIQ. For F5L, methylation values of the different CpG sites (beta values) were first logit-transformed and modelled as a linear function of the number of SNP alleles in a linear mixed regression model that included a random effect to account for the relatedness of the samples (implemented in GEMMA[60]); the model was furthermore adjusted for age, sex and cell-type proportions[61]. Significance was assessed using a LRT. For MARTHA, a linear regression model was used where the methylation values were taken as the outcome that was modelled using the allele dosage (from imputation) as the predictor, while adjusted for age, sex, batch, array, position of the sample on its array and measured cell-type compositions (lymphocytes, monocytes, polynuclear, eosinophils and basophils directly determined by ADVIA 120 Hematology System (Siemens Healthcare Diagnostics, Deerfield, IL)). Significance was assessed using a Wald test.

SNP and CpG site annotations. We annotated sets of genes or features (for example, non-coding RNAs) to SNPs by selecting the smallest genomic region that contained all SNPs in high LD with a SNP of interest (at $R^2>0.5$ according to 1000 Genomes[50] data or HapMap release 22 when the SNP is not indexed in 1000 Genomes Project) and annotated to that SNP all genes (that were extended by 5 kb in their promoter regions) or features falling at least partially in that window. We counted the number of features falling in regions flanking our meQTLs. By randomly selecting a similar number of SNPs among the ones that entered the analysis and generating LD-defined regions, we calculated an average absolute fold enrichment of genomic features (fold enrichment of features per Mb) based on 1,000 replicates. Significance of feature fold-enrichment was calculated as the proportion of random replicates (sets of random regions) having a number of features greater or equal to the observed number of features in our meQTL regions. These fold enrichments are not meant to be compared with what is expected in random regions of the genome *per se*, but rather random regions of the subset of the genome that is covered by our set of SNPs, as well as SNPs in LD with them.

A CpG site was annotated to a gene if its location falls from within 1,500 bases from the gene's transcription start sites up to and including its 3′untranslated region. This set of annotations is available with the HumanMethylation450 array documentation.

Functional interaction networks. We downloaded version 2013 of the FI network from Reactome (www.reactome.org) that we analysed using functions from the igraph package of R (www.r-project.org).

Expression data sets in CD4 and CD8 lymphocytes. The expression data from purified T-cell subpopulations was derived from multiple sclerosis (MS) patients ($n=68$) and healthy controls ($n=67$). MS cases were recruited through the Cambridge MS clinic and controls from the Cambridge (UK) BioResource (http://www.cambridgebioresource.org.uk). The study had approval from appropriate ethics committees and all subjects gave written informed consent. In total 50–80 ml of heparinized whole blood was collected from each individual and peripheral blood mononuclear cells (PBMCs) isolated using Ficoll density-gradient centrifugation. PBMCs were washed twice following Ficoll separation to remove platelet contamination. DNA was prepared from aliquots of the PBMCs using standard isolation protocols. CD4+ and CD8+ T cells were isolated using magnetic-activated cell sorting according to the manufacturers' instructions (Miltenyi Biotec). Specifically CD3+ cells were negatively selected using a Pan T Cell Isolation Kit followed by positive selection of CD8+ cells with the remaining fraction representing CD4+ cells. The purity of the separated cells was checked by flow cytometry for a subset of the samples.

The isolated cells were immediately lysed in TRIzol reagent (Life Technologies) and stored at $-80\,^{\circ}C$ prior to extraction. Total RNA was extracted according to the standard TRIzol protocol, and DNA contamination removed using DNase I treatment (Thermo Scientific). The extracted RNA was cleaned using the RNeasy MinElute Cleanup Kit (Qiagen), and the RNA integrity assessed using an Agilent 2100 Bioanalyzer and quantified using a Nanodrop 1000. Libraries for stranded

total RNA-sequencing were prepared using the Illumina Stranded Total RNA protocol (RS-122-2301). Libraries were assessed using the Agilent 2100 Bioanalyzer.

Samples were indexed and sequenced on Illumina HiSeq 2000 (paired-end 2×100 bp). Raw reads were trimmed for quality (phred33 ≥ 30), length ($n \geq 32$) and Illumina adapters using Trimmomatic v. 0.22 (ref. 62). Filtered reads were aligned to the hg19 human reference (for human samples) using Tophat v. 1.4.1 (ref. 63) and bowtie v. 0.12.8 (ref. 64). Duplicates were marked in the bam file using Picard's MarkDuplicates.jar v. 1.70 (http://picard.sourceforge.net). Raw read counts for ucsc and ensembl genes were obtained using htseq-count v. 0.5.3p9 (http://www-huber.embl.de/users/anders/HTSeq). FPKM values (fragments per kilobase of transcript per million fragments mapped) for ucsc and ensembl genes and transcripts were obtained using cufflinks v. 2.1.1 (ref. 65). To reduce the distorting effects of high data values in sequence data, we transformed all FPKM values with the inverse hyperbolic sine function $\text{asinh}\, x = \ln(x + \sqrt{(x2+1)})$ and finally performed quantile normalization across signal-transformed FPKMs. Genotyping on the 137 samples was performed on Illumina Human 2.5 M arrays, and additional variants were imputed (Impute2) using 1000 Genomes (phase 1 integrated version 3) UK sample as reference. Association testing for genotype versus normalized FPKMs was carried out using PLINK[66] using a linear framework for successfully imputed sites (info > 0.8) without covariates, combining the cases and controls but separating CD4+ and CD8+ to independent data sets. Significance was calculated using a Wald statistic (P_{Wald}).

References

1. Nica, A. C. *et al.* Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* **6,** e1000895 (2010).
2. Grundberg, E. *et al.* Population genomics in a disease targeted primary cell model. *Genome Res.* **19,** 1942–1952 (2009).
3. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27,** 72–79 (2011).
4. Westra, H. J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
5. Bell, J. T. *et al.* DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol.* **12,** R10 (2011).
6. Cotterchio, M. *et al.* Ontario familial colon cancer registry: methods and first-year response rates. *Chronic Dis. Can.* **21,** 81–86 (2000).
7. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39,** 989–994 (2007).
8. Chen, Y. A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–209 (2013).
9. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B* **1,** 289–300 (1995).
10. Shoemaker, R., Deng, J., Wang, W. & Zhang, K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. *Genome Res.* **20,** 883–889 (2010).
11. Smith, A. K. *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics* **15,** 145 (2014).
12. Heyn, H. *et al.* DNA methylation contributes to natural human variation. *Genome Res.* **23,** 1363–1372 (2013).
13. Hellman, A. & Chess, A. Extensive sequence-influenced DNA methylation polymorphism in the human genome. *Epigenetics Chromatin* **3,** 11 (2010).
14. van Eijk, K. R. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13,** 636 (2012).
15. Zhang, D. *et al.* Genetic control of individual differences in gene-specific methylation in human brain. *Am. J. Hum. Genet.* **86,** 411–419 (2010).
16. Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol.* **15,** R37 (2014).
17. Heyn, H. *et al.* Linkage of DNA methylation quantitative trait loci to human cancer risk. *Cell Rep.* **7,** 331–338 (2014).
18. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *Elife* **2,** e00523 (2013).
19. Shi, J. *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nat. Commun.* **27,** 3365 (2014).
20. Miremadi, A., Oestergaard, M. Z., Pharoah, P. D. & Caldas, C. Cancer genetics of epigenetic genes. *Hum. Mol. Genet.* **16**(Spec No 1): R28–R49 (2007).
21. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10,** 252–263 (2009).
22. Girard, A., Sachidanandam, R., Hannon, G. J. & Carmell, M. A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442,** 199–202 (2006).
23. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11,** R53 (2010).
24. Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.* **24,** 14–24 (2014).
25. Garvin, A. J. *et al.* The deSUMOylase SENP7 promotes chromatin relaxation for homologous recombination DNA repair. *EMBO Rep.* **14,** 975–983 (2013).
26. Tommerup, N. & Vissing, H. Isolation and fine mapping of 16 novel human zinc finger-encoding cDNAs identify putative candidate genes for developmental and malignant disorders. *Genomics* **27,** 259–264 (1995).
27. Sanchez-Vega, F. *et al.* Recurrent patterns of DNA methylation in the ZNF154, CASP8, and VHL promoters across a wide spectrum of human solid epithelial tumors and cancer cell lines. *Epigenetics* **8,** 1355–1372 (2013).
28. Okamoto, T. *et al.* Methylated-mediated repression of ZNF154 in ovarian cancer is associated with poor overall survival. *Cancer Res.* **72,** LB–87 (2012).
29. Arai, E. *et al.* Single-CpG-resolution methylome analysis identifies clinicopathologically aggressive CpG island methylator phenotype clear cell renal cell carcinomas. *Carcinogenesis* **33,** 1487–1493 (2012).
30. Reinert, T. *et al.* Comprehensive genome methylation analysis in bladder cancer: identification and validation of novel methylated genes and application of these as urinary tumor markers. *Clin. Cancer Res.* **1,** 5582–5592 (2011).
31. Reinert, T. *et al.* Diagnosis of bladder cancer recurrence based on urinary levels of EOMES, HOXA9, POU4F2, TWIST1, VIM, and ZNF154 hypermethylation. *PLoS ONE* **7,** e46297 (2012).
32. Rosenbloom, K. R. *et al.* ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic Acids Res.* **41,** D56–D63 (2013).
33. Meister, G. Argonaute proteins: functional insights and emerging roles. *Nat. Rev. Genet.* **14,** 447–459 (2013).
34. Bamezai, S., Rawat, V. P. & Buske, C. Concise review: the Piwi-piRNA axis: pivotal beyond transposon silencing. *Stem Cells* **30,** 2603–2611 (2012).
35. Popadin, K., Gutierrez-Arcelus, M., Dermitzakis, E. T. & Antonarakis, S. E. Genetic and epigenetic regulation of human lincRNA gene expression. *Am. J. Hum. Genet.* **93,** 1015–1026 (2013).
36. Huang, X. A. *et al.* A major epigenetic programming mechanism guided by piRNAs. *Dev. Cell* **24,** 502–516 (2013).
37. Yin, H. & Lin, H. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature* **450,** 304–308 (2007).
38. Gibbs, J. R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6,** e1000952 (2010).
39. Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3,** 462–474 (2012).
40. Grundberg, E. *et al.* Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am. J. Hum. Genet.* **93,** 876–890 (2013).
41. Lee, B. K. & Iyer, V. R. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. *J. Biol. Chem.* **287,** 30906–30913 (2012).
42. Bao, L., Zhou, M. & Cui, Y. CTCFBSDB: a CTCF-binding site database for characterization of vertebrate genomic insulators. *Nucleic Acids Res.* **36,** D83–D87 (2008).
43. Stadler, M. B. *et al.* DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* **480,** 490–495 (2011).
44. Szabó, P. E., Tang, S. H., Silva, F. J., Tsark, W. M. & Mann, J. R. Role of CTCF binding sites in the Igf2/H19 imprinting control region. *Mol. Cell Biol.* **24,** 4791–4800 (2004).
45. Gregor, A. *et al.* De novo mutations in the genome organizer CTCF cause intellectual disability. *Am. J. Hum. Genet.* **93,** 124–131 (2013).
46. Nikolic, T. *et al.* The DNA-binding factor Ctcf critically controls gene expression in macrophages. *Cell Mol. Immunol.* **11,** 58–70 (2014).
47. Newcomb, P. A. *et al.*Colon Cancer Family Registry. Colon Cancer Family Registry: an international resource for studies of the genetic epidemiology of colon cancer. *Cancer Epidemiol. Biomarkers Prev.* **16,** 2331–2343 (2007).
48. Triche, Jr T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41,** e90 (2013).
49. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29,** 189–196 (2013).
50. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
51. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31,** 142–147 (2013).
52. Koestler, D. C. *et al.* Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. *Cancer Epidemiol. Biomarkers Prev.* **21,** 1293–1302 (2012).
53. Accomando, W. P. *et al.* Decreased NK cells in patients with head and neck cancer determined in archival DNA. *Clin. Cancer Res.* **18,** 6147–6154 (2012).
54. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).

55. Gagnon, F., Aïssi, D., Carrié, A., Morange, P. E. & Trégouët, D. A. Robust validation of methylation levels association at CPT1A locus with lipid plasma levels. *J. Lipid Res.* **55,** 1189–1191 (2014).
56. Dick, K. J. *et al.* DNA methylation and body-mass index: a genome-wide analysis. *Lancet* **383,** 1990–1998 (2014).
57. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
58. Germain, M. *et al.* Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. *PLoS ONE* **7,** e38538 (2012).
59. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13,** R44 (2012).
60. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44,** 821–824 (2012).
61. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86 (2012).
62. Lohse, M. *et al.* RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res.* **40,** W622–W627 (2012).
63. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
64. Langmead, B., Trapnell, C., Prop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
65. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28,** 511–515 (2010).
66. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

Acknowledgements

OFCCR: we acknowledge the contribution of François Bacot and Sylvie LaBoissière of the McGill University and Génome Québec Innovation Centre, Montréal, Canada for the profiling of the HumanMethylation450 array in OFCCR. Funding sources for this study include grants from the Ontario Research Fund (GL2 competition), the Canadian Institutes of Health Research, European Community's Seventh Framework Programme—ENGAGE Consortium grant agreement HEALTH-F4-2007-201413. T.J.H. and B.W.Z. received Senior Investigator Awards from the Ontario Institute for Cancer Research, through generous support from the Ontario Ministry of Research and Innovation. OFCCR is a member of the Colon Cancer Family Registry (CCFR). CCFR is supported by the National Cancer Institute, National Institutes of Health under RFA # CA-95-011. OFCCR is supported by grant U01 CA074783.

MS: we acknowledge the use of subjects from the Cambridge BioResource and the support of the Cambridge NIHR Biomedical Research Centre. Funding sources include grants from the Cambridge NIHR Biomedical Research Centre and the UK Medical Research Council (G1100125). T.P. holds a Canada Research Chair and his contribution was supported by CIHR and FRSQ (RMGA).

MARTHA: funding sources include grants from the Program Hospitalier de Recherche Clinique and the GenMed LABEX (ANR-10-LABX-0013), the Canadian Institutes of Health Research (grant MOP 86466) and the Heart and Stroke Foundation of Canada (grant T6484). Statistical analyses were performed using the C2BIG computing cluster, funded by the Région Ile de France, Pierre and Marie Curie University, and the ICAN Institute for Cardiometabolism and Nutrition (ANR-10-IAHU-05). D.A. was supported by a PhD grant from the Région Ile de France (CORDDIM).

F5L: funding sources include grants from the Canadian Institutes of Health Research (MOP86466) and by the Heart and Stroke Foundation of Canada (T6484). F.G. holds a Canada Research Chair.

The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CFRs, nor does it mention of trade names, commercial products or organizations that imply endorsement by the US Government or the CFR.

Author contributions

B.W.Z., S.G. and T.J.H. designed the epigenetic study. M.L. and T.J.H. conceived the MeQTL study. M.B., T.P. and S.S. designed the expression study. P.-E.M., F.G. and P.S.W. contributed data. M.L., B.G., M.G., D.A. and I.K. performed statistical analysis under the supervision of T.J.H., T.P., D.-A.T. and F.G. S.H.E.Z., M.W. and T.J.H. performed case studies of specific loci. M.L., S.H.E.Z., T.P. and T.J.H. wrote the article. All authors read and approved the final manuscript.

Additional information

Accession codes. The OFCCR data were deposited in dbGaP under the accession number phs000779.v1.p1.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* 6:6326 doi: 10.1038/ncomms7326 (2015).

Quatrième partie

Conclusions & Perspectives

La méthylation de l'ADN est un phénomène de plus en plus étudié en génomique humaine, notamment pour essayer d'expliquer les mécanismes physiopathologiques associés aux pathologies complexes telles que les maladies cardiovasculaires. Lorsque j'ai débuté mon travail de thèse, il s'agissait d'un travail novateur de par le fait que la méthylation soit mesurée dans les cellules du sang périphérique ainsi que par le fait que les études épidémiologiques du méthylome étaient encore très peu développées. Mon travail de thèse dans lequel j'ai eu en charge l'ensemble des analyses bioinformatiques et biostatistiques des données de méthylation générées dans l'étude MARTHA, ainsi que la réplication des principaux résultats dans l'étude F5L-Pedigrees, a permis de montrer l'intérêt de l'étude de la méthylation de l'ADN à partir de cellules circulantes pour découvrir de nouveaux mécanismes biologiques. L'identification de la méthylation du gène *HIF3A* associée à l'indice de masse corporelle (Dick *et al.* 2014) a été réalisée sur un peu plus de 3 600 individus alors que l'implication de ce locus n'avait pas été suspectée par l'approche GWAS sur 100 000 individus (Speliotes *et al.* 2010), ni même sur 340 000 individus (Locke *et al.* 2015). Il en est de même pour l'identification de la méthylation du gène *CPT1A* associée aux taux plasmatiques de lipides (Gagnon *et al.* 2014) obtenue avec un peu plus de 500 individus. Le rôle de *CPT1A* dans le métabolisme des lipides n'a pas été décelé par une méthode GWAS sur 190 000 individus (Global Lipids Genetics Consortium 2013). Enfin, le travail sur l'identification des sites CpG dont la méthylation est sous l'influence de mSNPs auquel j'ai contribué (Lemire *et al.* 2015) permet de générer de nouvelles hypothèses concernant des mécanismes d'interaction entre des gènes. Ces travaux montrent clairement que les méthodes étudiant le méthylome (MWAS) sont complémentaires de celles étudiant la variabilité génétique du génome (GWAS). Les associations trouvées dans ce travail de thèse sont le fruit d'analyses statistiques et nécessiteraient désormais d'être validées par d'autres méthodes mesurant la méthylation (par exemple via du séquençage) ainsi que par des études expérimentales fonctionnelles afin de caractériser finement les mécanismes biologiques qui se cachent derrière elles.

En outre, il est nécessaire d'avoir recours à de grandes études épidémiologiques pour pouvoir détecter des associations entre les niveaux de méthylation mesurés à partir de cellules sanguines et un phénotype d'intérêt car les effets sont plus petit que dans les tissus plus pertinents. Cela a été montré dans le travail qui a permis d'identifier la méthylation du gène *HIF3A* comme associé à l'IMC. L'association trouvée dans le sang est plus faible que celle trouvée dans le tissu adipeux. La méthylation de l'ADN étant un phénomène tissu-spécifique, il est donc très important de bien sélectionner le tissu dans lequel on souhaite mesurer la méthylation. La mauvaise sélection du tissu

pourrait aboutir à ne pas détecter d'association entre les niveaux de méthylation et le phénotype d'intérêt.

Au cours de mon travail de thèse, j'ai également étudié d'autres approches statistiques et stratégies d'analyse pour optimiser l'analyse des données de méthylation que j'avais à ma disposition. Si je n'ai pu obtenir de résultats probants, certaines pistes méritent d'être poursuivies. Dans les paragraphes suivants, j'expose certains points dont l'étude, je pense, mériterait d'être poursuivie.

6 Autres approches statistiques

Normalisation & Correction : Au début de ce travail de thèse, la puce HM450k venait d'être mise au point et peu de littérature sur le sujet était disponible. Les biais de la puce n'étaient pas encore tous connus et bien caractérisés. Il existait donc peu de méthodes de correction. Depuis, de nombreuses équipes ont travaillé sur le sujet, les biais étant plus étudiés, et de nouvelles méthodes de correction et normalisation sont publiées fréquemment. Bien que les méthodes de correction utilisées dans ce travail de thèse soient celles reconnues comme étant parmi les plus performantes, il est possible que de nouvelles méthodes (par exemple "FunNorm" de Fortin *et al.* 2014 ou la dernière version non publiée de la méthode proposée par Touleimat & Tost 2012) puissent améliorer la qualité des données produites et puissent permettre de trouver des signaux plus faibles et non détectés via les méthodes de correction utilisées dans ce travail de thèse.

Régression Beta : La mesure du niveau de méthylation est représentée par une valeur β comprise entre 0 et 1 dont la distribution suit parfois une loi bêta. Il est donc tout à fait naturel d'appliquer des modèles de régression bêta (Saadati & Benner 2014 ; Wahl *et al.* 2014) lorsque l'on s'intéresse aux niveaux de méthylation comme variables à expliquer. Néanmoins, les conclusions des différentes études qui se sont intéressées à ce type de modèle de régression divergent quant à la supériorité de ces modèles dans ce contexte.

Régression Quantile : Les modèles de régression quantile (Bind *et al.* 2015 ; Briollais & Durrieu 2014) ne cherchent pas à modéliser l'espérance (c'est-à-dire la moyenne) d'une variable d'intérêt mais un quantile prédéfini. Il est possible par exemple d'étudier la médiane (quantile 50%) qui serait moins influencée par les données aberrantes que la moyenne. L'exemple le plus connu d'utilisation de régression par quantile est la courbe d'évolution du poids des nourrissons dans les carnets de santé

français (figure 5.3). Sur la figure 5.3 est représentée l'évolution du poids des nourrissons selon les quantiles 3%, 25%, 75% et 97%. En d'autres termes, les courbes représentent l'évolution du poids : des 3% des nourrissons ayant le poids le plus important (courbe pour le quantile 97%) ; des 25% des nourrissons ayant le poids le plus important (courbe pour le quantile 75%) ; des 25% des nourrissons ayant le poids le plus faible (courbe pour le quantile 25%) ; des 3% des nourrissons ayant le poids le plus faible (courbe pour le quantile 3%).

Ces régressions permettent d'obtenir non pas l'évolution de la moyenne comme pour les régressions linéaires simples mais d'obtenir l'évolution de la distribution de la variable d'intérêt, elles permettent donc une description beaucoup plus précise.

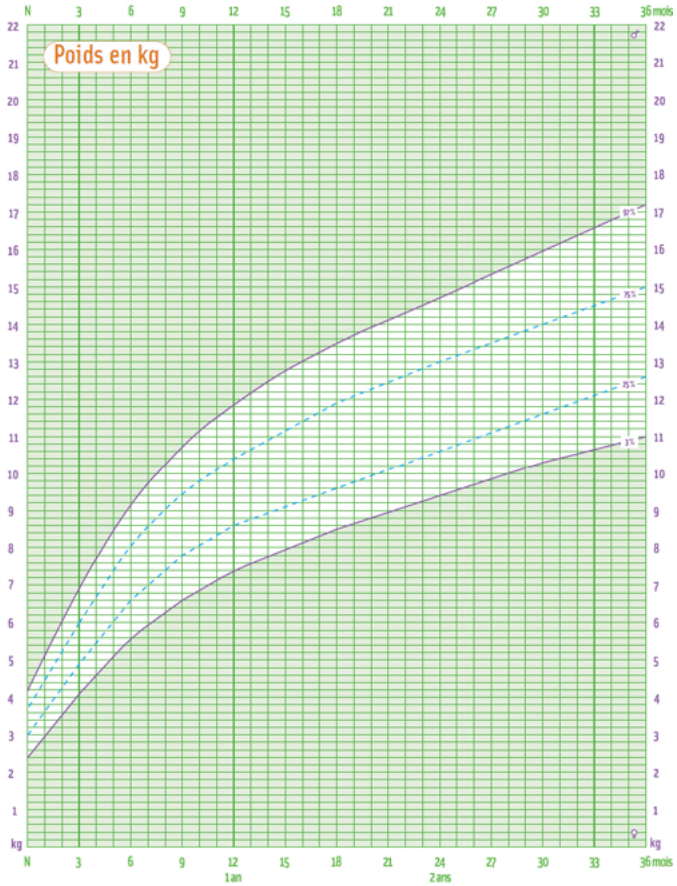

FIGURE 5.3 – Exemple de régression quantile non linéaire : la courbe d'évolution du poids des nourrissons en fonction de l'âge dans les carnets de santé français.

On peut imaginer que des profils de méthylation particuliers peuvent non pas influencer la moyenne du trait étudié mais plutôt d'autres paramètres comme des quantiles spécifiques ; quantiles potentiellement pathogènes. J'ai utilisé les régressions quantile sur plusieurs variables de l'hémostase (facteur de von Willebrand, facteur VIII) pour essayer d'augmenter la puissance de détection de sites CpG

associés à ces variables. Malheureusement, les résultats que j'ai obtenus, quel que soit le quantile étudié, n'ont pas apporté d'information nouvelle par rapport aux modèles de régression linéaires classiques. Cela ne signifie pas que l'application de ce type de modèle ne puisse pas, pour d'autres traits, être plus puissante pour identifier des marques de méthylation. Leur étude mériterait de plus amples investigations.

Analyse des sites CpG par région : Différentes méthodes statistiques ont été proposées pour étudier simultanément l'effet de plusieurs sites CpG adjacents. Cela pourrait permettre de détecter des effets plus modestes de la méthylation mais sur une zone plus large. L'inconvénient de ces méthodes est qu'elles ne prennent pas en compte la corrélation entre les sites CpG. Nous pouvons citer : Bump Hunting (Jaffe, Murakami *et al.* 2012), Comb-p (Pedersen *et al.* 2012), A-clustering (Sofer *et al.* 2013), Probe Lasso (Butcher & S. Beck 2015), GAMP (Zhao *et al.* 2015) ou encore DMRcate (Peters *et al.* 2015). J'ai également essayé plusieurs de ces méthodes sur différentes variables mais les résultats n'étaient pas non plus suffisamment concluants.

Analyse des sites CpG par cluster : Il est également possible de considérer non pas les sites CpG proches mais des clusters de sites CpG co-méthylés (éloignés ou non) en utilisant la méthode WGCNA (Langfelder & Horvath 2008 ; B. Zhang & Horvath 2005) ou la méthode MICA (Rau *et al.* 2013). La méthode WGCNA consiste à créer un réseau de sites CpG dont les niveaux de méthylation sont corrélés entre eux permettant ainsi de découvrir des gènes appartenant à une même voie métabolique (Spiers *et al.* 2015). L'application de cette stratégie aux données de MARTHA mériterait d'être envisagée. La principale limite que j'ai pu identifier pour le moment est liée à la "lourdeur" des calculs associés si l'on désirait estimer les corrélations entre plus de 300 000 variables. La première étape de cette stratégie est de calculer la matrice de corrélation entre les 388 120 variables des niveaux de méthylation. Les calculs via R se déroulent en mémoire vive, or la matrice de corrélation pèse environ 500 gigaoctets de mémoire sachant qu'un noeud du cluster ne possède que 64 gigaoctets de mémoire vive, il n'est donc pas aisé de réaliser cette étape. La réalisation de cette première étape nécessite le développement d'une stratégie permettant de s'affranchir de la limite des capacités informatiques à notre disposition.

7 Autres approches biologiques

Analyse de la méthylation dans un autre type cellulaire : La méthylation étant un phénomène type cellulaire spécifique, il est tout à fait possible que des mécanismes de méthylation impliqués dans la régulation de certains traits complexes ne soient pas détectables à partir de l'ADN des cellules sanguines. Il peut être intéressant de mesurer la méthylation sur d'autres types cellulaires plus pertinents pour la pathologie étudiée, comme par exemple les cellules cardiaques, le foie, le tissu adipeux, etc (Roadmap Epigenomics Consortium *et al.* 2015).

Analyse de la méthylation à haute densité : Un des points inconvénients de la puce HumanMethylation450k est qu'elle ne mesure la méthylation que sur une petite portion (moins de 2%) des sites CpG du génome (Bibikova, Barnes *et al.* 2011). Il est donc raisonnable de penser que nous ayons pu passer à côté de sites CpG dont la méthylation est associée à la variable d'intérêt. Une méthode peu coûteuse déjà appliquée aux polymorphismes est l'imputation qui permet d'estimer statistiquement les polymorphismes non mesurés grâce au phénomène de déséquilibre de liaison (corrélation entre des polymorphismes proches). Ces méthodes commencent à émerger pour imputer les niveaux de méthylation pour les sites CpG non mesurés (Ernst & Kellis 2015) notamment en ce qui concerne les données de la puce HM450k (W. Zhang *et al.* 2015). Par le fait que ces méthodes soient récentes et pas encore optimisées, je ne les ai pas testées au cours de mon travail de thèse.

Un autre type de méthode est d'utiliser le séquençage à haut débit pour mesurer la méthylation de l'ensemble des sites CpG du génome et ainsi avoir une représentation globale du méthylome. Malgré l'avantage certain de ce type de méthode, il possède un inconvénient : le coût est beaucoup plus élevé que de mesurer la méthylation avec la puce HumanMethylation450k. Le prix pour mesurer le niveau de méthylation d'un individu avec la puce HM450k est d'environ 285 € (prix proposé par le TCAG de Toronto lors de la rédaction de cette thèse et comprenant la puce ainsi que la main d'œuvre), alors que pour mesurer le niveau de méthylation d'environ 3,7 millions de site CpGs avec du Methyl-Seq (via le kit Agilent SureSelectXT Human) il est de 1300 €. Le tarif peut même atteindre 50000 € pour du *Whole-Genome Bisulfite Sequencing*.

Analyse de l'hydroxyméthylation : Depuis peu, un certain intérêt est porté aux cytosines hydroxyméthylées (Kriaucionis & Heintz 2009 ; Pfeifer & Szabo 2009). L'hydroxyméthylation est l'ajout d'un groupement hydroxyle (OH) à la suite du groupement méthyle (CH3). Les cytosines hy-

droxyméthylées possèdent donc un groupement hydroxyméthyle (CH2OH). Des recherches montrent que l'hydroxyméthylation est impliquée dans certains cancers (Pfeifer, Kadam *et al.* 2013 ; Rodger *et al.* 2014 ; Uribe-Lewis *et al.* 2015), dans des pathologies neurodégénératives (Al-Mahdawi *et al.* 2014 ; Massart *et al.* 2014) ou dans la régulation transcriptomique (Branco *et al.* 2012), cela ouvre donc la voie à de nouvelles hypothèses.

Le protocole standard de traitement au bisulfite de sodium de l'ADN ne permet pas de distinguer les cytosines méthylées (5-mC) des cytosines hydroxyméthylées (5-hmC). Un nouveau traitement a été mis au point permettant de distinguer ces deux formes de méthylation des cytosines : le traitement oxydant au bisulfite de sodium (Ito *et al.* 2011 ; Tahiliani *et al.* 2009). Avec le traitement standard, les cytosines méthylées et hydroxyméthylées sont protégées de la désamination du bisulfite de sodium et ne sont donc pas converties en uraciles. Tandis qu'avec le nouveau traitement, une étape d'oxydation par la protéine TET est ajoutée avant le traitement au bisulfite de sodium ce qui va transformer le groupement hydroxyle des cytosines hydroxyméthylées en un groupement formyle pour les convertir en cytosines formylées (5-fC). Le bisulfite de sodium va ensuite convertir les cytosines non méthylées et formylées en uraciles puis l'amplification va convertir les uraciles en thymines. Il suffit donc de comparer les résultats obtenus par le traitement au bisulfite de sodium et de ceux obtenus par le traitement oxydant au bisulfite de sodium pour connaître l'état de la cytosine : non méthylée, méthylée ou hydroxyméthylée (figure 5.4).

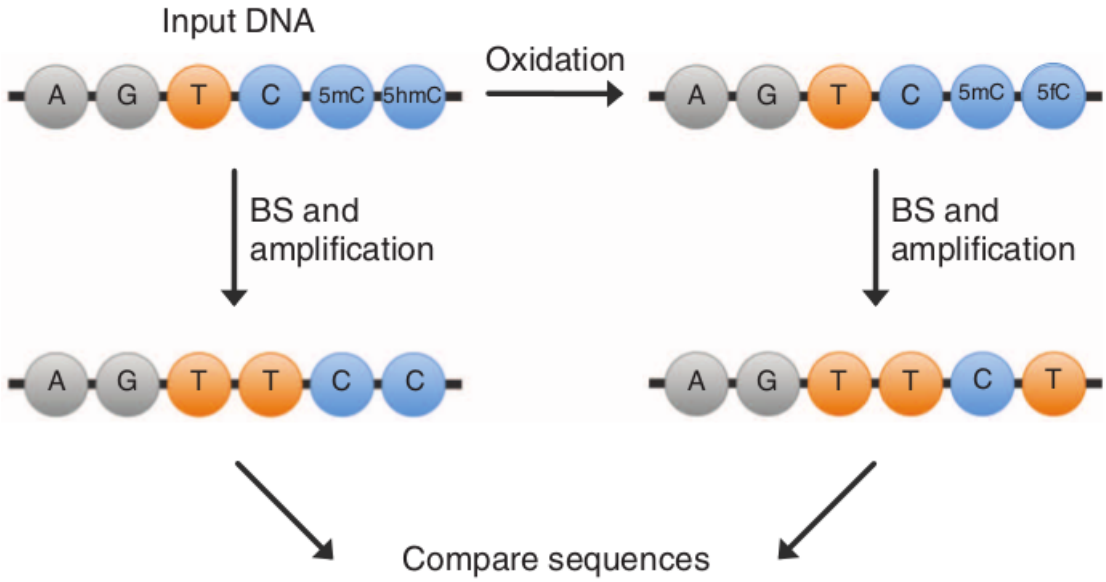

Figure 5.4 – Schéma de la transformation des cytosines en fonction de leur état : non méthylé, méthylé ou hydroxyméthylé. Extrait de Booth, Branco *et al.* 2012

Ce protocole est déjà appliqué à différentes méthodes de séquençage, comme le TAB-Seq pour *Tet-assisted bisulfite sequencing* (Rusk 2012 ; Yu *et al.* 2012) ou le oxBS-Seq pour *oxidative bisulfite sequencing* (Booth, Branco *et al.* 2012 ; Booth, Ost *et al.* 2013), mais aussi à la puce Illumina HumanMethylation450k (Field *et al.* 2015 ; Nazor *et al.* 2014 ; Stewart *et al.* 2015). Toutes ces méthodes intègrent l'étape d'oxydation permettant de convertir les cytosines hydroxyméthylées en cytosines formylées et ainsi les différencier des cytosines méthylées. Le prix pour mesurer l'hydroxyméthylation avec la puce HM450k serait d'au moins le double du prix pour mesurer seulement la méthylation soit environ 640$ par individu. Dans un premier temps, il faudrait mesurer le niveau de méthylation de l'individu via une première puce HM450k et un traitement au bisulfite de sodium puis dans un second temps et via une seconde puce HML450k, il faudrait mesurer les niveaux d'hydroxyméthylation avec un traitement d'oxydation puis au bisulfite de sodium.

Il est possible d'imaginer que d'autres modifications épigénétiques des cytosines (comme la 5-formylcytosine (5-fC), la 5-carboxylcytosine (5-caC) ou la 3-méthylcytosine (3-mC), figure 5.5), dans le cas où leur stabilité dans le génome est confirmée, peuvent également influencer certains mécanismes biologiques. Des travaux allant dans ce sens sont déjà publiés notamment en ce qui concerne la 5-carboxylcytosine (Inoue *et al.* 2011), la 5-formylcytosine (Inoue *et al.* 2011 ; Z. Sun, N. Dai *et al.* 2015) et la 3-méthylcytosine (Camps & Eichman 2011).

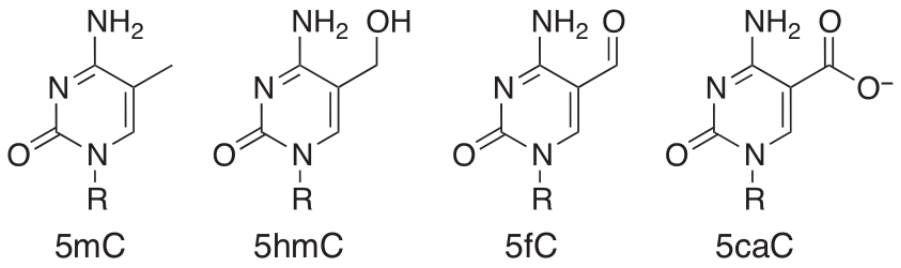

Figure 5.5 – Exemples de modification épigénétique des cytosines. Extrait de Booth, Ost *et al.* 2013

Analyse des autres modifications épigénétiques : Enfin il est également possible d'étudier d'autres modifications épigénétiques. Certaines peuvent concerner l'ARN tandis que d'autres concernent les protéines comme par exemple les histones. Les histones sont des petites protéines permettant la condensation de l'ADN et dont leurs lysines [1] peuvent avoir des marques de méthylation (Martin & Y. Zhang 2005) et/ou d'acétylation (Bannister & Kouzarides 2011 ; Cedar & Bergman 2009 ;

1. Les protéines sont constituées d'acides aminés dont notamment la lysine.

Zhou *et al.* 2011). Les protéines peuvent également avoir des marques épigénétiques dont notamment la phosphorylation (ajout d'un groupement phosphate PO_4^{3-}), l'ubiquitination (ajout d'une protéine ubiquitine), la sumoylation (ajout d'une protéine SUMO), etc. Les ARNs quant à eux peuvent avoir des marques épigénétiques comme les ARNs messager dont l'adénine peut être méthylée en position N6 (Meyer *et al.* 2012).

De par la récente émergence de l'étude du méthylome à l'échelle épidémiologique et le large éventail de perspectives présentées ci-dessus, l'épidémiologie épigénétique n'en est qu'à ses débuts et a un bel avenir devant elle promettant de nombreuses découvertes permettant de mieux comprendre les mécanismes de régulation du génome.

Bibliographie

1. Aïssi, D. *et al.* Genome-Wide Investigation of DNA Methylation Marks Associated with FV Leiden Mutation. *PLoS ONE* **9,** e108087 (29 sept. 2014).

2. Altman, D. G. & Bland, J. M. Measurement in Medicine : The Analysis of Method Comparison Studies. *Journal of the Royal Statistical Society. Series D (The Statistician)* **32,** 307–317. issn : 0039-0526 (1er sept. 1983).

3. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proceedings of the National Academy of Sciences* **90,** 11995–11999. issn : 0027-8424, 1091-6490 (15 déc. 1993).

4. Antoni, G. *et al.* A multi-stage multi-design strategy provides strong evidence that the BAI3 locus is associated with early-onset venous thromboembolism. *Journal of Thrombosis and Haemostasis* **8,** 2671–2679. issn : 1538-7836 (2010).

5. Aryee, M. J. *et al.* Minfi : a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369. issn : 1367-4803, 1460-2059 (15 mai 2014).

6. Baccarelli, A., Rienstra, M. & Benjamin, E. J. Cardiovascular Epigenetics Basic Concepts and Results From Animal and Human Studies. *Circulation : Cardiovascular Genetics* **3,** 567–573. issn : 1942-325X, 1942-3268 (1er déc. 2010).

7. Bannister, A. J. & Kouzarides, T. Regulation of chromatin by histone modifications. *Cell Research* **21,** 381–395. issn : 1001-0602 (mar. 2011).

8. Beck, S. Taking the measure of the methylome. *Nature Biotechnology* **28,** 1026–1028. issn : 1087-0156 (oct. 2010).

9. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. GWAS Central : a comprehensive resource for the comparison and interrogation of genome-wide association studies. *European Journal of Human Genetics* **22,** 949–952. issn : 1018-4813 (2014).

10. Bedford, M. T. & Richard, S. Arginine Methylation : An Emerging Regulator of Protein Function. *Molecular Cell* **18,** 263–272. issn : 1097-2765 (2005).

11. Bell, C. G., Finer, S. *et al.* Integrated Genetic and Epigenetic Analysis Identifies Haplotype-Specific Methylation in the FTO Type 2 Diabetes and Obesity Susceptibility Locus. *PLoS ONE* **5,** e14040 (18 nov. 2010).

12. Bell, C. G., Teschendorff, A. E. *et al.* Genome-wide DNA methylation analysis for diabetic nephropathy in type 1 diabetes mellitus. *BMC Medical Genomics* **3,** 33. issn : 1755-8794 (5 août 2010).

13. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300. issn : 0035-9246 (1er jan. 1995).

14. Beshlawi, I. *et al.* Thiamine responsive megaloblastic anemia : The puzzling phenotype. *Pediatric Blood & Cancer* **61,** 528–531. issn : 1545-5017 (1er mar. 2014).

15. Bhat, T. *et al.* Neutrophil to lymphocyte ratio and cardiovascular diseases : a review. *Expert Review of Cardiovascular Therapy* **11,** 55–59. issn : 1477-9072 (1er jan. 2013).

16. Bibikova, M., Barnes, B. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics. New Genomic Technologies and Applications* **98,** 288–295. issn : 0888-7543 (oct. 2011).

17. Bibikova, M., Le, J. *et al.* Genome-wide DNA methylation profiling using Infinium® assay. *Epigenomics* **1,** 177–200. issn : 1750-1911 (1er oct. 2009).

18. Bind, M.-A. C. *et al.* Beyond the Mean : Quantile Regression to Explore the Association of Air Pollution with Gene-Specific Methylation in the Normative Aging Study. *Environmental Health Perspectives.* issn : 1552-9924. doi :10.1289/ehp.1307824 (13 mar. 2015).

19. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Research* **8,** 1499–1504. issn : 0305-1048, 1362-4962 (11 avr. 1980).

20. Bjornsson, H. T., Daniele Fallin, M. & Feinberg, A. P. An integrated epigenetic and genetic approach to common human disease. *Trends in Genetics* **20,** 350–358. issn : 0168-9525 (2004).

21. Booth, M. J., Branco, M. R. *et al.* Quantitative Sequencing of 5-Methylcytosine and 5-Hydroxymethylcytosine at Single-Base Resolution. *Science* **336,** 934–937. issn : 0036-8075, 1095-9203 (18 mai 2012).

22. Booth, M. J., Ost, T. W. B. *et al.* Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. *Nature Protocols* **8,** 1841–1851. issn : 1754-2189 (2013).

23. Bourc'his, D., Xu, G.-L., Lin, C.-S., Bollman, B. & Bestor, T. H. Dnmt3L and the Establishment of Maternal Genomic Imprints. *Science* **294,** 2536–2539. issn : 0036-8075, 1095-9203 (21 déc. 2001).

24. Branco, M. R., Ficz, G. & Reik, W. Uncovering the role of 5-hydroxymethylcytosine in the epigenome. *Nature Reviews Genetics* **13,** 7–13. issn : 1471-0056 (jan. 2012).

25. Brandeis, M. *et al.* Spl elements protect a CpG island from de novo methylation. *Nature* **371,** 435–438 (29 sept. 1994).

26. Breitling, L. P., Yang, R., Korn, B., Burwinkel, B. & Brenner, H. Tobacco-Smoking-Related Differential DNA Methylation : 27K Discovery and Replication. *The American Journal of Human Genetics* **88,** 450–457. issn : 0002-9297 (2011).

27. Brenner, C. *et al.* Myc represses transcription through recruitment of DNA methyltransferase corepressor. *The EMBO Journal* **24,** 336–346. issn : 0261-4189, 1460-2075 (26 jan. 2005).

28. Brinkman, A. B. *et al.* Whole-genome DNA methylation profiling using MethylCap-seq. *Methods. DNA Methylation Analysis* **52,** 232–236. issn : 1046-2023 (nov. 2010).

29. Briollais, L. & Durrieu, G. Application of quantile regression to recent genetic and -omic studies. *Human Genetics* **133,** 951–966. issn : 0340-6717, 1432-1203 (26 avr. 2014).

30. Burton, A. Hypermethylation of HIC1 causes cancer. *The Lancet Oncology* **4,** 133. issn : 1470-2045 (mar. 2003).

31. Butcher, L. M. & Beck, S. Probe Lasso : A novel method to rope in differentially methylated regions with 450K DNA methylation data. *Methods. (Epi)Genomics approaches and their applications* **72,** 21–28. issn : 1046-2023 (15 jan. 2015).

32. Camps, M. & Eichman, B. F. Unraveling a Connection between DNA Demethylation Repair and Cancer. *Molecular Cell* **44,** 343–344. issn : 1097-2765 (4 nov. 2011).

33. Cedar, H. & Bergman, Y. Linking DNA methylation and histone modification : patterns and paradigms. *Nature Reviews Genetics* **10,** 295–304. issn : 1471-0056 (2009).

34. Chang, H.-C., Cho, C.-Y. & Hung, W.-C. Downregulation of RECK by promoter methylation correlates with lymph node metastasis in non-small cell lung cancer. *Cancer Science* **98,** 169–173. issn : 1349-7006 (2007).

35. Chen, Y.-a. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8,** 203–209. issn : 1559-2294 (2013).

36. Chen, W. Y. *et al.* Heterozygous disruption of Hic1 predisposes mice to a gender-dependent spectrum of malignant tumors. *Nature Genetics* **33,** 197–202. issn : 1061-4036 (2003).

37. Christensen, B. C. *et al.* Aging and Environmental Exposures Alter Tissue-Specific DNA Methylation Dependent upon CpG Island Context. *PLoS Genet* **5,** e1000602 (2009).

38. Consortium, T. 1. G. P. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073. issn : 0028-0836 (28 oct. 2010).

39. Consortium, T. H. M. P. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214. issn : 0028-0836 (2012).

40. Córdova-Palomera, A. *et al.* Genome-wide methylation study on depression : differential methylation and variable methylation in monozygotic twins. *Translational Psychiatry* **5,** e557 (2015).

41. Dai, W. *et al.* Methylation Linear Discriminant Analysis (MLDA) for identifying differentially methylated CpG islands. *BMC Bioinformatics* **9,** 337. issn : 1471-2105 (8 août 2008).

42. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3,** 771–784. issn : 1750-1911 (25 nov. 2011).

43. Demerath, E. W. *et al.* Epigenome-wide Association Study (EWAS) of BMI, BMI Change, and Waist Circumference in African American Adults Identifies Multiple Replicated Loci. *Human Molecular Genetics,* ddv161. issn : 0964-6906, 1460-2083 (1er mai 2015).

44. Diaz, G. A., Banikazemi, M., Oishi, K., Desnick, R. J. & Gelb, B. D. Mutations in a new gene encoding a thiamine transporter cause thiamine-responsive megaloblastic anaemia syndrome. *Nature Genetics* **22,** 309–312. issn : 1061-4036 (1999).

45. Dick, K. J. *et al.* DNA methylation and body-mass index : a genome-wide analysis. *The Lancet* **383,** 1990–1998. issn : 0140-6736 (2014).

46. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587. issn : 1471-2105 (30 nov. 2010).

47. Du, Q., Luu, P.-L., Stirzaker, C. & Clark, S. J. Methyl-CpG-binding domain proteins : readers of the epigenome. *Epigenomics,* 1–23. issn : 1750-192X (30 avr. 2015).

48. Dudoit, S., Shaffer, J. P. & Boldrick, J. C. Multiple Hypothesis Testing in Microarray Experiments. *Statistical Science* **18,** 71–103. issn : 0883-4237, 2168-8745 (fév. 2003).

49. Eijk, K. R. v. *et al.* Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics* **13,** 636. issn : 1471-2164 (17 nov. 2012).

50. Ernst, J. & Kellis, M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nature Biotechnology* **advance online publication.** issn : 1087-0156. doi :10.1038/nbt.3157. <http://www.nature.com/nbt/journal/vaop/ncurrent/full/nbt.3157.html> (visité le 27/03/2015) (2015).

51. Esteller, M. Cancer epigenomics : DNA methylomes and histone-modification maps. *Nature Reviews Genetics* **8,** 286–298. issn : 1471-0056 (2007).

52. Esteller, M. Epigenetics in Cancer. *New England Journal of Medicine* **358,** 1148–1159. issn : 0028-4793 (13 mar. 2008).

53. Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447,** 433–440. issn : 0028-0836 (2007).

54. Feinberg, A. P. & Tycko, B. The history of cancer epigenetics. *Nature Reviews Cancer* **4,** 143–153. issn : 1474-175X (2004).

55. Field, S. F. *et al.* Accurate Measurement of 5-Methylcytosine and 5-Hydroxymethylcytosine in Human Cerebellum DNA by Oxidative Bisulfite on an Array (OxBS-Array). *PLoS ONE* **10,** e0118202 (2015).

56. Fisher, R. A. *Statistical methods for research workers.* Fourth ed. - revised and enlarged. xiii, 307 (Oliver & Boyd, Edinburgh, 1932).

57. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biology* **15,** 503. issn : 1465-6906 (3 déc. 2014).

58. Frazier-Wood, A. C. *et al.* Methylation at CPT1A locus is associated with lipoprotein subfraction profiles. *Journal of Lipid Research* **55,** 1324–1330. issn : 0022-2275, 1539-7262 (1er juil. 2014).

59. Gagnon, F., Aïssi, D., Carrié, A., Morange, P.-E. & Trégouët, D.-A. Robust validation of methylation levels association at CPT1A locus with lipid plasma levels. *Journal of Lipid Research* **55,** 1189–1191. issn : 0022-2275, 1539-7262 (1er juil. 2014).

60. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13,** 539–552. issn : 1465-4644, 1468-4357 (1er juil. 2012).

61. Garnier, S. *et al.* Genome-Wide Haplotype Analysis of Cis Expression Quantitative Trait Loci in Monocytes. *PLoS Genet* **9,** e1003240 (31 jan. 2013).

62. Gebhard, C. *et al.* General Transcription Factor Binding at CpG Islands in Normal Cells Correlates with Resistance to De novo DNA Methylation in Cancer Cells. *Cancer Research* **70,** 1398–1407. issn : 0008-5472, 1538-7445 (15 fév. 2010).

63. Global Lipids Genetics Consortium. Discovery and refinement of loci associated with lipid levels. *Nature Genetics* **45,** 1274–1283. issn : 1061-4036 (nov. 2013).

64. Greenwood, E. Gender-dependent tumour suppression. *Nature Reviews Cancer* **3,** 86–86. issn : 1474-175X (2003).

65. Grundberg, E. *et al.* Global Analysis of DNA Methylation Variation in Adipose Tissue from Twins Reveals Links to Disease-Associated Variants in Distal Regulatory Elements. *The American Journal of Human Genetics* **93,** 876–890. issn : 0002-9297 (7 nov. 2013).

66. Gu, H. *et al.* Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nature Protocols* **6,** 468–481. issn : 1754-2189 (2011).

67. Guh, D. P. *et al.* The incidence of co-morbidities related to obesity and overweight : A systematic review and meta-analysis. *BMC Public Health* **9,** 88. issn : 1471-2458 (25 mar. 2009).

68. Guintivano, J., Aryee, M. J. & Kaminsky, Z. A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8,** 290–302. issn : 1559-2294 (1er mar. 2013).

69. Gutierrez-Arcelus, M. *et al.* Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife* **2,** e00523. issn : 2050-084X (4 juin 2013).

70. Heinig, M. *et al.* A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature* **467,** 460–464. issn : 0028-0836 (23 sept. 2010).

71. Hendrich, B., Hardeland, U., Ng, H. H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature* **401,** 301–304. issn : 0028-0836 (16 sept. 1999).

72. Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends in Genetics* **19,** 269–277. issn : 0168-9525 (2003).

73. Herceg, Z. Epigenetics and cancer : towards an evaluation of the impact of environmental and dietary factors. *Mutagenesis* **22,** 91–103. issn : 0267-8357, 1464-3804 (1er mar. 2007).

74. Herceg, Z. & Hainaut, P. Genetic and epigenetic alterations as biomarkers for cancer detection, diagnosis and prognosis. *Molecular Oncology* **1,** 26–41. issn : 1574-7891 (2007).

75. Herceg, Z. & Vaissière, T. Epigenetic mechanisms and cancer : An interface between the environment and the genome. *Epigenetics* **6,** 804–819. issn : 1559-2294 (2011).

76. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine* **21,** 1539–1558. issn : 1097-0258 (2002).

77. Hill, V. K. *et al.* Genome-Wide DNA Methylation Profiling of CpG Islands in Breast Cancer Identifies Novel Genes Associated with Tumorigenicity. *Cancer Research* **71,** 2988–2999. issn : 0008-5472, 1538-7445 (15 avr. 2011).

78. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Research* **22,** 271–282. issn : 1088-9051, 1549-5469 (1er fév. 2012).

79. Horvath, S. *et al.* Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proceedings of the National Academy of Sciences* **103,** 17402–17407. issn : 0027-8424, 1091-6490 (14 nov. 2006).

80. Houseman, E. A., Kelsey, K. T., Wiencke, J. K. & Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data : a mathematical perspective. *BMC Bioinformatics* **16,** 95. issn : 1471-2105 (21 mar. 2015).

81. Houseman, E. A., Accomando, W. P. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86. issn : 1471-2105 (8 mai 2012).

82. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30,** 1431–1439. issn : 1367-4803, 1460-2059 (15 mai 2014).

83. Hsiung, D. T. *et al.* Global DNA Methylation Level in Whole Blood as a Biomarker in Head and Neck Squamous Cell Carcinoma. *Cancer Epidemiology Biomarkers & Prevention* **16,** 108–114. issn : 1055-9965, 1538-7755 (1er jan. 2007).

84. Huang, T. T.-K. *et al.* Mobilisation of public support for policy actions to prevent obesity. *The Lancet*. issn : 0140-6736. doi :10.1016/S0140-6736(14)61743-8. <http://www.sciencedirect.com/science/article/pii/S0140673614617438> (visité le 27/03/2015) (2015).

85. Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F. & Botella, J. Assessing heterogeneity in meta-analysis : Q statistic or I2 index ? *Psychological Methods* **11,** 193–206. issn : 1082-989X (juin 2006).

86. Huynh, J. L. *et al.* Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. *Nature Neuroscience* **17,** 121–130. issn : 1097-6256 (jan. 2014).

87. Illingworth, R. S. & Bird, A. P. CpG islands – 'A rough guide'. *FEBS Letters. Prague Special Issue : Functional Genomics and Proteomics* **583,** 1713–1720. issn : 0014-5793 (2009).

88. Inoue, A., Shen, L., Dai, Q., He, C. & Zhang, Y. Generation and replication-dependent dilution of 5fC and 5caC during mouse preimplantation development. *Cell Research* **21,** 1670–1676. issn : 1001-0602 (2011).

89. Ioannidis, J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in Meta-Analyses of Genome-Wide Association Investigations. *PLoS ONE* **2,** e841 (5 sept. 2007).

90. Irizarry, R. A., Hobbs, B. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264. issn : 1465-4644, 1468-4357 (1er avr. 2003).

91. Irizarry, R. A., Ladd-Acosta, C. *et al.* Comprehensive high-throughput arrays for relative methylation (CHARM). *Genome Research* **18,** 780–790. issn : 1088-9051, 1549-5469 (1er mai 2008).

92. Irvin, M. R. *et al.* Epigenome-Wide Association Study of Fasting Blood Lipids in the Genetics of Lipid-Lowering Drugs and Diet Network Study. *Circulation* **130,** 565–572. issn : 0009-7322, 1524-4539 (12 août 2014).

93. Ito, S. *et al.* Tet Proteins Can Convert 5-Methylcytosine to 5-Formylcytosine and 5-Carboxylcytosine. *Science* **333,** 1300–1303. issn : 0036-8075, 1095-9203 (2 sept. 2011).

94. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology* **15,** R31. issn : 1465-6906 (4 fév. 2014).

95. Jaffe, A. E., Murakami, P. *et al.* Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International Journal of Epidemiology* **41,** 200–209. issn : 0300-5771, 1464-3685 (1er fév. 2012).

96. Januar, V., Saffery, R. & Ryan, J. Epigenetics and depressive disorders : a review of current progress and future directions. *International Journal of Epidemiology,* dyu273. issn : 0300-5771, 1464-3685 (24 fév. 2015).

97. Ji, H. *et al.* Comprehensive methylome map of lineage commitment from haematopoietic progenitors. *Nature* **467,** 338–342. issn : 0028-0836 (16 sept. 2010).

98. Ji, Y., Wu, C., Liu, P., Wang, J. & Coombes, K. R. Applications of beta-mixture models in bioinformatics. *Bioinformatics* **21,** 2118–2122. issn : 1367-4803, 1460-2059 (1er mai 2005).

99. Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V. & Jordan, I. K. On the presence and role of human gene-body DNA methylation. *Oncotarget* **3,** 462–474. issn : 1949-2553 (9 mai 2012).

100. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8,** 118–127. issn : 1465-4644, 1468-4357 (1er jan. 2007).

101. Jones, P. A. & Baylin, S. B. The Epigenomics of Cancer. *Cell* **128,** 683–692. issn : 0092-8674 (2007).

102. Joubert, B. R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives* **120,** 1425–1431. issn : 1552-9924 (oct. 2012).

103. Kaminen-Ahola, N. *et al.* Maternal Ethanol Consumption Alters the Epigenotype and the Phenotype of Offspring in a Mouse Model. *PLoS Genet* **6,** e1000811 (15 jan. 2010).

104. Kile, M. L. *et al.* Effect of prenatal arsenic exposure on DNA methylation and leukocyte subpopulations in cord blood. *Epigenetics* **9,** 774–782. issn : 1559-2294 (2014).

105. Kim, J. *et al.* DNA Methylation in Inflammatory Genes among Children with Obstructive Sleep Apnea. *American Journal of Respiratory and Critical Care Medicine* **185,** 330–338. issn : 1073-449X (2012).

106. Koestler, D. C., Avissar-Whiting, M., Houseman, E. A., Karagas, M. R. & Marsit, C. J. Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. *Environmental Health Perspectives* **121,** 971–977. issn : 1552-9924 (août 2013).

107. Koestler, D. C., Christensen, B. C. *et al.* Blood-based profiles of DNA methylation predict the underlying distribution of cell types. *Epigenetics* **8,** 816–826. issn : 1559-2294 (2013).

108. Kriaucionis, S. & Heintz, N. The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain. *Science (New York, N.Y.)* **324,** 929–930. issn : 1095-9203 (15 mai 2009).

109. Kuan, P. F., Wang, S., Zhou, X. & Chu, H. A statistical framework for Illumina DNA methylation arrays. *Bioinformatics* **26,** 2849–2855. issn : 1367-4803, 1460-2059 (15 nov. 2010).

110. Kuasne, H. *et al.* Genome-wide methylation and transcriptome analysis in penile carcinoma : uncovering new molecular markers. *Clinical Epigenetics* **7,** 46. issn : 1868-7083 (18 avr. 2015).

111. Laird, P. W. Principles and challenges of genome-wide DNA methylation analysis. *Nature Reviews Genetics* **11,** 191–203. issn : 1471-0056 (2010).

112. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921. issn : 0028-0836 (2001).

113. Langfelder, P. & Horvath, S. WGCNA : an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559. issn : 1471-2105 (29 déc. 2008).

114. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nature Communications* **6.** doi :10.1038/ncomms7326. <http://www.nature.com/ncomms/2015/150226/ncomms7326/full/ncomms7326.html> (visité le 26/03/2015) (2015).

115. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366,** 362–365 (1993).

116. Liang, L. *et al.* An epigenome-wide association study of total serum immunoglobulin E concentration. *Nature* **520,** 670–674. issn : 0028-0836 (2015).

117. Lienert, F. *et al.* Identification of genetic elements that autonomously determine DNA methylation states. *Nature Genetics* **43,** 1091–1097. issn : 1061-4036 (nov. 2011).

118. Lin, L. I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* **45,** 255–268. issn : 0006-341X (1er mar. 1989).

119. Lister, R. *et al.* Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* **462,** 315–322. issn : 0028-0836 (19 nov. 2009).

120. Liu, G., Yang, F., Han, B., Liu, J. & Nie, G. Identification of four SLC19A2 mutations in four Chinese thiamine responsive megaloblastic anemia patients without diabetes. *Blood Cells, Molecules, and Diseases* **52,** 203–204. issn : 1079-9796 (2014).

121. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology* **31,** 142–147. issn : 1087-0156 (2013).

122. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206. issn : 0028-0836 (2015).

123. Long, H. K., Blackledge, N. P. & Klose, R. J. ZF-CxxC domain-containing proteins, CpG islands and the chromatin connection. *Biochemical Society Transactions* **41,** 727–740. issn : 0300-5127 (Pt 3 1er juin 2013).

124. Long, N. K. *et al.* Hypermethylation of the RECK gene predicts poor prognosis in oral squamous cell carcinomas. *Oral Oncology* **44,** 1052–1058. issn : 1368-8375 (nov. 2008).

125. Mägi, R. & Morris, A. P. GWAMA : software for genome-wide association meta-analysis. *BMC Bioinformatics* **11,** 288. issn : 1471-2105 (28 mai 2010).

126. Al-Mahdawi, S., Virmouni, S. A. & Pook, M. A. The emerging role of 5-hydroxymethylcytosine in neurodegenerative diseases. *Neurogenomics* **8,** 397 (2014).

127. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN : Subset-quantile Within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips. *Genome Biology* **13,** R44. issn : 1465-6906 (15 juin 2012).

128. Martin, C. & Zhang, Y. The diverse functions of histone lysine methylation. *Nature Reviews Molecular Cell Biology* **6,** 838–849. issn : 1471-0072 (nov. 2005).

129. Massart, R. *et al.* The genome-wide landscape of DNA methylation and hydroxymethylation in response to sleep deprivation impacts on synaptic plasticity genes. *Translational Psychiatry* **4,** e347 (21 jan. 2014).

130. Matouk, C. C. & Marsden, P. A. Epigenetic Regulation of Vascular Endothelial Gene Expression. *Circulation Research* **102,** 873–887. issn : 0009-7330, 1524-4571 (25 avr. 2008).

131. Meyer, K. D. *et al.* Comprehensive Analysis of mRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell* **149,** 1635–1646. issn : 0092-8674 (22 juin 2012).

132. Michels, K. B. *et al.* Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods* **10,** 949–955. issn : 1548-7091 (2013).

133. Mikstiene, V. *et al.* Thiamine responsive megaloblastic anemia syndrome : A novel homozygous SLC19A2 gene mutation identified. *American Journal of Medical Genetics Part A,* n/a–n/a. issn : 1552-4833 (2015).

134. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* **16,** 1215–1215. issn : 0305-1048, 1362-4962 (11 fév. 1988).

135. Miller-Delaney, S. F. C. *et al.* Differential DNA methylation profiles of coding and non-coding genes define hippocampal sclerosis in human temporal lobe epilepsy. *Brain,* awu373. issn : 0006-8950, 1460-2156 (30 déc. 2014).

136. Miranda, T. B. & Jones, P. A. DNA methylation : The nuts and bolts of repression. *Journal of Cellular Physiology* **213,** 384–390. issn : 1097-4652 (1er nov. 2007).

137. Moore, L. E. *et al.* Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the Spanish Bladder Cancer Study : a case–control study. *The Lancet Oncology* **9,** 359–366. issn : 1470-2045 (2008).

138. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38,** 23–38. issn : 0893-133X (jan. 2013).

139. Morris, K. V., Chan, S. W.-L., Jacobsen, S. E. & Looney, D. J. Small Interfering RNA-Induced Transcriptional Gene Silencing in Human Cells. *Science* **305,** 1289–1292. issn : 0036-8075, 1095-9203 (27 août 2004).

140. Movassagh, M. *et al.* Differential DNA Methylation Correlates with Differential Expression of Angiogenic Factors in Human Heart Failure. *PLoS ONE* **5,** e8564 (13 jan. 2010).

141. Nazor, K. L. *et al.* Application of a low cost array-based technique — TAB-Array — for quantifying and mapping both 5mC and 5hmC at single base resolution in human pluripotent stem cells. *Genomics. 5-hydroxymethylation* **104,** 358–367. issn : 0888-7543 (nov. 2014).

142. Nica, A. C. *et al.* The Architecture of Gene Regulatory Variation across Multiple Human Tissues : The MuTHER Study. *PLoS Genet* **7,** e1002003 (2011).

143. Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252,** 1097–1102. issn : 0036-8075, 1095-9203 (24 mai 1991).

144. Ogden, C. L., Carroll, M. D., Kit, B. K. & Flegal, K. M. Prevalence of childhood and adult obesity in the United States, 2011-2012. *JAMA* **311,** 806–814. issn : 1538-3598 (26 fév. 2014).

145. Okano, M., Bell, D. W., Haber, D. A. & Li, E. DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. *Cell* **99,** 247–257. issn : 0092-8674 (29 oct. 1999).

146. Oudot-Mellakh, T. *et al.* Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein C anticoagulant pathway : the MARTHA project. *British Journal of Haematology* **157,** 230–239. issn : 1365-2141 (2012).

147. Pan, H. *et al.* HIF3A association with adiposity : the story begins before birth. *Epigenomics,* 1–13. issn : 1750-192X (26 mai 2015).

148. Paul, D. S. & Beck, S. Advances in epigenome-wide association studies for common diseases. *Trends in Molecular Medicine* **20,** 541–543. issn : 1471-4914 (oct. 2014).

149. Pedersen, B. S., Schwartz, D. A., Yang, I. V. & Kechris, K. J. Comb-p : software for combining, analyzing, grouping and correcting spatially correlated P-values. *Bioinformatics* **28,** 2986–2988. issn : 1367-4803, 1460-2059 (15 nov. 2012).

150. Peters, T. J. *et al.* De novo identification of differentially methylated regions in the human genome. *Epigenetics & Chromatin* **8,** 6. issn : 1756-8935 (27 jan. 2015).

151. Pfeifer, G. P., Kadam, S. & Jin, S.-G. 5-hydroxymethylcytosine and its potential roles in development and cancer. *Epigenetics & Chromatin* **6,** 10. issn : 1756-8935 (1er mai 2013).

152. Pfeifer, G. P. & Szabo, P. E. 5-hydroxymethylcytosine, a modified mammalian DNA base with a potential regulatory role. *Epigenomics* **1,** 21–22. issn : 1750-192X (oct. 2009).

153. Pfeiffer, L. *et al.* DNA Methylation of Lipid-Related Genes Affects Blood Lipid Levels. *Circulation : Cardiovascular Genetics* **8,** 334–342. issn : 1942-325X, 1942-3268 (1er avr. 2015).

154. Philibert, R., Plume, J. M., Gibbons, F. X., Brody, G. H. & Beach, S. The impact of recent alcohol use on genome wide DNA methylation signatures. *Behavioral and Psychiatric Genetics* **3,** 54 (2012).

155. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14,** 293. issn : 1471-2164 (1er mai 2013).

156. Pollard, K. S., Dudoit, S. & Laan, M. J. v. d. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (éds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 249–271 (Springer New York, 2005). isbn : 978-0-387-25146-2 978-0-387-29362-2. <http://link.springer.com/chapter/10.1007/0-387-29362-0_15> (visité le 24/03/2015).

157. Rakyan, V. K., Beyan, H. *et al.* Identification of Type 1 Diabetes–Associated DNA Methylation Variable Positions That Precede Disease Diagnosis. *PLoS Genet* **7,** e1002300 (29 sept. 2011).

158. Rakyan, V. K., Down, T. A., Balding, D. J. & Beck, S. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* **12,** 529–541. issn : 1471-0056 (1er août 2011).

159. Rasool, M. *et al.* The role of epigenetics in personalized medicine : challenges and opportunities. *BMC Medical Genomics* **8,** S5. issn : 1755-8794 (Suppl 1 15 jan. 2015).

160. Rau, C. D. *et al.* Maximal information component analysis : a novel non-linear network analysis method. *Statistical Genetics and Methodology* **4,** 28 (2013).

161. Reinius, L. E. *et al.* Differential DNA Methylation in Purified Human Blood Cells : Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE* **7,** e41361 (2012).

162. Richardson, B. Primer : epigenetics of autoimmunity. *Nature Reviews Rheumatology* **3,** 521–527. issn : 1759-4790 (1er sept. 2007).

163. Riggs, A. D. & Jones, P. A. 5-methylcytosine, gene regulation, and cancer. *Advances in Cancer Research* **40,** 1–30. issn : 0065-230X (1983).

164. Ritchie, M. E. *et al.* A comparison of background correction methods for two-colour microarrays. *Bioinformatics* **23,** 2700–2707. issn : 1367-4803, 1460-2059 (15 oct. 2007).

165. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330. issn : 0028-0836 (2015).

166. Robertson, K. D. DNA methylation and chromatin – unraveling the tangled web. *, Published online : 05 August 2002; | doi :10.1038/sj.onc.1205609* **21.** doi :10.1038/sj.onc.1205609. <http://www.nature.com/onc/journal/v21/n35/full/1205609a.html> (visité le 24/04/2015) (5 août 2002).

167. Robertson, K. D. DNA methylation and human disease. *Nature Reviews Genetics* **6,** 597–610. issn : 1471-0056 (2005).

168. Rocanin-Arjo, A. *et al.* A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential. *Blood* **123,** 777–785. issn : 0006-4971, 1528-0020 (30 jan. 2014).

169. Rocañín-Arjó, A. *et al.* Thrombin Generation Potential and Whole-Blood DNA methylation. *Thrombosis Research* **135,** 561–564. issn : 0049-3848 (mar. 2015).

170. Rocke, D. M. & Durbin, B. Approximate variance-stabilizing transformations for gene-expression microarray data. *Bioinformatics* **19,** 966–972. issn : 1367-4803, 1460-2059 (22 mai 2003).

171. Rodger, E. J., Chatterjee, A. & Morison, I. M. 5-hydroxymethylcytosine : a potential therapeutic target in cancer. *Epigenomics* **6,** 503–514. issn : 1750-192X (2014).

172. Rönn, T. *et al.* Impact of age, BMI and HbA1c levels on the genome-wide DNA methylation and mRNA expression patterns in human adipose tissue and identification of epigenetic biomarkers in blood. *Human Molecular Genetics,* ddv124. issn : 0964-6906, 1460-2083 (10 avr. 2015).

173. Rotival, M. *et al.* Integrating Genome-Wide Genetic Variations and Monocyte Expression Data Reveals Trans-Regulated Gene Modules in Humans. *PLoS Genet* **7,** e1002367 (2011).

174. Rusk, N. Epigenetics : The sixth base and counting. *Nature Methods* **9,** 646–646. issn : 1548-7091 (2012).

175. Saadati, M. & Benner, A. Statistical challenges of high-dimensional methylation data. *Statistics in Medicine* **33,** 5347–5357. issn : 1097-0258 (2014).

176. Sabatine, M. S. *et al.* Metabolomic Identification of Novel Biomarkers of Myocardial Ischemia. *Circulation* **112,** 3868–3875. issn : 0009-7322, 1524-4539 (20 déc. 2005).

177. Sabbah, C., Mazo, G., Paccard, C., Reyal, F. & Hupé, P. SMETHILLIUM : spatial normalization METHod for ILLumina InfinIUM HumanMethylation BeadChip. *Bioinformatics* **27,** 1693–1695. issn : 1367-4803, 1460-2059 (15 juin 2011).

178. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6,** 692–702. issn : 1559-2294 (2011).

179. Schübeler, D. Function and information content of DNA methylation. *Nature* **517,** 321–326. issn : 0028-0836 (15 jan. 2015).

180. Serre, D., Lee, B. H. & Ting, A. H. MBD-isolated Genome Sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. *Nucleic Acids Research* **38,** 391–399. issn : 0305-1048, 1362-4962 (1er jan. 2010).

181. Siegmund, K. D. Statistical approaches for the analysis of DNA methylation microarray data. *Human Genetics* **129,** 585–595. issn : 0340-6717, 1432-1203 (26 avr. 2011).

182. Siegmund, K. D. & Laird, P. W. Analysis of complex methylation data. *Methods* **27,** 170–178. issn : 1046-2023 (2002).

183. Søes, S. *et al.* Hypomethylation and increased expression of the putative oncogene ELMO3 are associated with lung cancer development and metastases formation. *Oncoscience* **1,** 367–374. issn : 2331-4737 (23 mai 2014).

184. Sofer, T., Schifano, E. D., Hoppin, J. A., Hou, L. & Baccarelli, A. A. A-clustering : a novel method for the detection of co-regulated methylation regions, and regions associated with exposure. *Bioinformatics* **29,** 2884–2891. issn : 1367-4803, 1460-2059 (15 nov. 2013).

185. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics* **42,** 937–948. issn : 1061-4036 (nov. 2010).

186. Spiers, H. *et al.* Methylomic trajectories across human fetal brain development. *Genome Research* **25,** 338–352. issn : 1549-5469 (mar. 2015).

187. Sreekumar, A. *et al.* Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature* **457,** 910–914. issn : 0028-0836 (2009).

188. Stewart, S. K. *et al.* oxBS-450K : A method for analysing hydroxymethylation using 450K BeadChips. *Methods. (Epi)Genomics approaches and their applications* **72,** 9–15. issn : 1046-2023 (15 jan. 2015).

189. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100,** 9440–9445. issn : 0027-8424, 1091-6490 (5 août 2003).

190. Stringhini, S. *et al.* Life-course socioeconomic status and DNA methylation of genes regulating inflammation. *International Journal of Epidemiology,* dyv060. issn : 0300-5771, 1464-3685 (17 avr. 2015).

191. Sun, Y. V. *et al.* Comparison of the DNA methylation profiles of human peripheral blood cells and transformed B-lymphocytes. *Human Genetics* **127,** 651–658. issn : 0340-6717, 1432-1203 (18 mar. 2010).

192. Sun, Z., Chai, H. S. *et al.* Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics* **4,** 84. issn : 1755-8794 (16 déc. 2011).

193. Sun, Z., Dai, N. *et al.* A Sensitive Approach to Map Genome-wide 5-Hydroxymethylcytosine and 5-Formylcytosine at Single-Base Resolution. *Molecular Cell* **57,** 750–761. issn : 1097-2765 (2015).

194. Tahiliani, M. *et al.* Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science* **324,** 930–935. issn : 0036-8075, 1095-9203 (15 mai 2009).

195. Tahir, S. *et al.* A novel homozygous SLC19A2 mutation in a Portuguese patient with diabetes mellitus and thiamine-responsive megaloblastic anaemia. *International Journal of Pediatric Endocrinology* **2015,** 6. issn : 1687-9856 (15 avr. 2015).

196. Talens, R. P. *et al.* Variation, patterns, and temporal stability of DNA methylation : considerations for epigenetic epidemiology. *The FASEB Journal* **24,** 3135–3144. issn : 0892-6638, 1530-6860 (1er sept. 2010).

197. Teschendorff, A. E., Marabita, F. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29,** 189–196. issn : 1367-4803, 1460-2059 (15 jan. 2013).

198. Teschendorff, A. E., Menon, U. *et al.* An Epigenetic Signature in Peripheral Blood Predicts Active Ovarian Cancer. *PLoS ONE* **4,** e8274 (2009).

199. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713. issn : 0028-0836 (2010).

200. Tobi, E. W. *et al.* DNA methylation signatures link prenatal famine exposure to growth and metabolism. *Nature Communications* **5.** doi :10.1038/ncomms6592. <http://www.nature.com/ncomms/2014/141126/ncomms6592/full/ncomms6592.html> (visité le 04/05/2015) (26 nov. 2014).

201. in (éd. Tost, J.) *Methods in Molecular Biology* 507 (Humana Press, 2009). isbn : 978-1-934115-61-9 978-1-59745-522-0. <http://link.springer.com/protocol/10.1007/978-1-59745-522-0_1> (visité le 26/05/2015).

202. Tost, J. & Gut, I. G. DNA methylation analysis by pyrosequencing. *Nature Protocols* **2,** 2265–2275. issn : 1754-2189 (2007).

203. Touleimat, N. & Tost, J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4,** 325–341. issn : 1750-1911 (2012).

204. Trégouët, D.-A. *et al.* Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk : results from a GWAS approach. *Blood* **113,** 5298–5303. issn : 0006-4971, 1528-0020 (21 mai 2009).

205. Triche, T. J., Weisenberger, D. J., Berg, D. V. D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Research* **41,** e90–e90. issn : 0305-1048, 1362-4962 (1er avr. 2013).

206. Tsai, P.-C. & Bell, J. T. Power and sample size estimation for epigenome-wide association scans to detect differential DNA methylation. *International Journal of Epidemiology,* dyv041. issn : 0300-5771, 1464-3685 (13 mai 2015).

207. Tsaprouni, L. G. *et al.* Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics* **9,** 1382–1396. issn : 1559-2294 (3 oct. 2014).

208. Turnbaugh, P. J., Hamady, M. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484. issn : 0028-0836 (22 jan. 2009).

209. Turnbaugh, P. J., Ley, R. E. *et al.* An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444,** 1027–131. issn : 0028-0836 (2006).

210. Uribe-Lewis, S. *et al.* 5-hydroxymethylcytosine marks promoters in colon that resist DNA hypermethylation in cancer. *Genome Biology* **16,** 69. issn : 1465-6906 (1er avr. 2015).

211. Vineis, P. *et al.* DNA methylation changes associated with cancer risk factors and blood levels of vitamin metabolites in a prospective study. *Epigenetics* **6,** 195–201. issn : 1559-2294 (2011).

212. Waalwijk, C. & Flavell, R. A. MspI, an isoschizomer of hpaII which cleaves both unmethylated and methylated hpaII sites. *Nucleic Acids Research* **5,** 3231–3236. issn : 0305-1048 (sept. 1978).

213. Wahl, S. *et al.* On the potential of models for location and scale for genome-wide DNA methylation data. *BMC Bioinformatics* **15,** 232. issn : 1471-2105 (3 juil. 2014).

214. Walsh, C. P., Chaillet, J. R. & Bestor, T. H. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature Genetics* **20,** 116–117. issn : 1061-4036 (oct. 1998).

215. Weber, M. *et al.* Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. *Nature Genetics* **37,** 853–862. issn : 1061-4036 (2005).

216. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research* **42,** D1001–D1006. issn : 0305-1048, 1362-4962 (D1 1er jan. 2014).

217. Wichmann, H.-E., Gieger, C., Illig, T. & MONICA/KORA Study Group. KORA-gen–resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen (Bundesverband Der Ärzte Des Öffentlichen Gesundheitsdienstes (Germany))* **67 Suppl 1,** S26–30. issn : 0941-3790 (août 2005).

218. Wierda, R. J., Geutskens, S. B., Jukema, J. W., Quax, P. H. & van den Elsen, P. J. Epigenetics in atherosclerosis and inflammation. *Journal of Cellular and Molecular Medicine* **14,** 1225–1240. issn : 1582-4934 (2010).

219. Winkelmann, J. *et al.* Mutations in DNMT1 cause autosomal dominant cerebellar ataxia, deafness and narcolepsy. *Human Molecular Genetics* **21,** 2205–2210. issn : 0964-6906, 1460-2083 (15 mai 2012).

220. Withers, B. E. & Dunbar, J. C. The endonuclease isoschizomers, SmaI and XmaI, bend DNA in opposite orientations. *Nucleic Acids Research* **21,** 2571–2577. issn : 0305-1048 (11 juin 1993).

221. Wolff, E. M. *et al.* Hypomethylation of a LINE-1 Promoter Activates an Alternate Transcript of the MET Oncogene in Bladders with Cancer. *PLoS Genet* **6,** e1000917 (2010).

222. Yokochi, T. & Robertson, K. D. Preferential Methylation of Unmethylated DNA by Mammaliande Novo DNA Methyltransferase Dnmt3a. *Journal of Biological Chemistry* **277,** 11735–11745. issn : 0021-9258, 1083-351X (5 avr. 2002).

223. Yu, M. *et al.* Base-Resolution Analysis of 5-Hydroxymethylcytosine in the Mammalian Genome. *Cell* **149,** 1368–1380. issn : 0092-8674 (2012).

224. Zeilinger, S. *et al.* Tobacco Smoking Leads to Extensive Genome-Wide Changes in DNA Methylation. *PLoS ONE* **8,** e63812 (2013).

225. Zhang, B. & Horvath, S. A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Genetics and Molecular Biology* **4.** <http://www.degruyter.com/view/j/sagmb.2005.4.1/sagmb.2005.4.1.1128/sagmb.2005.4.1.1128.xml;jsessionid=B09F36FBD4F74826EB9E6BB7748EA4A0> (visité le 25/03/2015) (12 août 2005).

226. Zhang, R. *et al.* Genome-wide DNA methylation analysis in alcohol dependence. *Addiction Biology* **18,** 392–403. issn : 1369-1600 (1er mar. 2013).

227. Zhang, W., Spector, T. D., Deloukas, P., Bell, J. T. & Engelhardt, B. E. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biology* **16,** 14. issn : 1465-6906 (24 jan. 2015).

228. Zhang, X., Mu, W. & Zhang, W. On the analysis of the Illumina 450K array data : probes ambiguously mapped to the human genome. *Behavioral and Psychiatric Genetics,* 73 (2012).

229. Zhao, N. *et al.* Global Analysis of Methylation Profiles From High Resolution CpG Data. *Genetic Epidemiology* **39,** 53–64. issn : 1098-2272 (2015).

230. Zhou, V. W., Goren, A. & Bernstein, B. E. Charting histone modifications and the functional organization of mammalian genomes. *Nature Reviews Genetics* **12,** 7–18. issn : 1471-0056 (jan. 2011).

231. Zhuang, J., Widschwendter, M. & Teschendorff, A. E. A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics* **13,** 59. issn : 1471-2105 (24 avr. 2012).

232. Zou, J., Lippert, C., Heckerman, D., Aryee, M. & Listgarten, J. Epigenome-wide association studies without the need for cell-type composition. *Nature Methods* **11,** 309–311. issn : 1548-7091 (2014).

Annexes

Articles

— **Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism** ; Germain M., Chasman D.I., de Haan H., Tang W., Lindström S., Weng L.-C., de Andrade M., de Visser M.C.H., Wiggins K.L., Suchon P., Saut N., Smadja D.M., Le Gal G., van Hylckama Vlieg A., Di Narzo A., Hao K., Nelson C.P, Rocanin-Arjo A., Folkersen L., Monajemi R., Rose L.M., Brody J.A., Slagboom E., **Aïssi D.**, Gagnon F., Deleuze J.F., Deloukas P., Tzourio C., Dartigues J.F, Berr C., Taylor K.D, Civelek M., Eriksson P., Cardiogenics Consortium, Psaty B.M., Houwing-Duitermaat J., Goodall A.H., Cambien F., Kraft P., Amouyel P., Samani N.J., Basu S., Ridker P.M., Rosendaal F.R., Kabrhel C., Folsom A.R., Heit J., Reitsma P.H, Trégouët D.A., Smith N.L., Morange P.E. ; *The American Journal of Human Genetics* 2015, DOI: 10.1016/j.ajhg.2015.01.019

ARTICLE

Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism

Marine Germain[1,2,3], Daniel I. Chasman[4], Hugoline de Haan[5], Weihong Tang[6], Sara Lindström[7], Lu-Chen Weng[6], Mariza de Andrade[8], Marieke C.H. de Visser[9], Kerri L. Wiggins[10], Pierre Suchon[11,12,13], Noémie Saut[11,12,13], David M. Smadja[14,15,16], Grégoire Le Gal[17,18], Astrid van Hylckama Vlieg[5], Antonio Di Narzo[19], Ke Hao[19], Christopher P. Nelson[20,21], Ares Rocanin-Arjo[1,2,3], Lasse Folkersen[22], Ramin Monajemi[23], Lynda M. Rose[24], Jennifer A. Brody[25], Eline Slagboom[26], Dylan Aïssi[1,2,3], France Gagnon[27], Jean-Francois Deleuze[28], Panos Deloukas[29,30], Christophe Tzourio[31], Jean-Francois Dartigues[31], Claudine Berr[32], Kent D. Taylor[33], Mete Civelek[34], Per Eriksson[35], Cardiogenics Consortium, Bruce M. Psaty[25,36], Jeanine Houwing-Duitermaat[23], Alison H. Goodall[20,21], François Cambien[1,2,3], Peter Kraft[7], Philippe Amouyel[37,38], Nilesh J. Samani[20,21], Saonli Basu[39], Paul M. Ridker[4], Frits R. Rosendaal[5], Christopher Kabrhel[40], Aaron R. Folsom[6], John Heit[41], Pieter H. Reitsma[9], David-Alexandre Trégouët[1,2,3], Nicholas L. Smith[10,36,42] and Pierre-Emmanuel Morange[11,12,13]*

*Correspondence:
pierre.morange@ap-hm.fr
[11] Laboratory of Haematology, La Timone Hospital, 13385 Marseille, France
[12] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, 13385 Marseille, France
[13] Nutrition Obesity and Risk of Thrombosis, Aix-Marseille University, UMR_S 1062, 13385 Marseille, France
Full list of author information is available at the end of the article

NOTE:
This file is an Author's Post-print version using the BioMed Central TeX template.
For the Publisher's Version/PDF, please see :
The American Journal of Human Genetics, 2015 Apr 2;96(4):532-42.
DOI: 10.1016/j.ajhg.2015.01.019.

Summary

Venous thromboembolism (VTE), the third leading cause of cardiovascular mortality, is a complex thrombotic disorder with environmental and genetic determinants. Although several genetic variants have been found associated with VTE, they explain a minor proportion of VTE risk in cases. We undertook a meta-analysis of genome-wide association studies (GWASs) to identify additional VTE susceptibility genes. Twelve GWASs totaling 7,507 VTE case subjects and 52,632 control subjects formed our discovery stage where 6,751,884 SNPs were tested for association with VTE. Nine loci reached the genome-wide significance level of $5 * 10^{-8}$ including six already known to associate with VTE (*ABO*, *F2*, *F5*, *F11*, *FGG*, and *PROCR*) and three unsuspected loci. SNPs mapping to these latter were selected for replication in three independent case-control studies totaling 3,009 VTE-affected individuals and 2,586 control subjects. This strategy led to the identification and replication of two VTE-associated loci, *TSPAN15* and *SLC44A2*, with lead risk alleles associated with odds ratio for disease of 1.31 ($p = 1.67 * 10^{-16}$) and 1.21 ($p = 2.75 * 10^{-15}$), respectively. The lead SNP at the *TSPAN15* locus is the intronic rs78707713 and the lead *SLC44A2* SNP is the non-synonymous rs2288904 previously shown to associate with transfusion-related acute lung injury. We further showed that these

two variants did not associate with known hemostatic plasma markers. *TSPAN15* and *SLC44A2* do not belong to conventional pathways for thrombosis and have not been associated to other cardiovascular diseases nor related quantitative biomarkers. Our findings uncovered unexpected actors of VTE etiology and pave the way for novel mechanistic concepts of VTE pathophysiology.

Introduction

Venous thromboembolism (VTE [MIM 188050]) is a common multicausal thrombotic disease with an annual incidence of 1 per 1,000. It includes two main clinical manifestations: deep vein thrombosis (DVT) and pulmonary embolism (PE). The latter is associated with a 1-year mortality of 20%, making VTE the third leading cause of cardiovascular death in industrialized countries [1]. Moreover, among survivors, 25%–50% will have lasting debilitating health problems such as post-thrombotic syndrome, severely hampering mobility and quality of life.

Factors contributing to VTE include endothelial injury or activation, reduced blood flow, and hypercoagulability of the blood, the so-called Virchow triad [2]. Venous thromboembolism has a strong genetic basis that is characterized by an underlying heritability estimate of 50% and a risk of developing the disease in an individual with an affected sib 2.5 higher than for the general population [3, 4]. But, like other complex phenotypes, most genetic contributors have not been elucidated because the proportion of heritability explained by replicated variants has been small [5, 6].

There are seven well-established genetic risk factors for VTE, all responsible for inherited hypercoagulable states. The first three are heterozygous deficiencies of the natural coagulation inhibitors (antithrombin, protein C, and protein S). These deficiencies are relatively rare, affecting <1% of the general population, and they increase VTE risk by approximately ten. The other four, Factor V (FV [MIM 612309]) Leiden, prothrombin (MIM 176930) G20210A, fibrinogen γ' (FGG) (MIM 134850) rs2066865, and blood group non-O, are more frequent with prevalence in European-descent individuals around 5% for the former two and ∼ 25% for the latter two. The increase in VTE risk is about 3-fold for the FV Leiden (RefSeq accession number NP_000121.2, p.Arg534Gln [c.1601G>A]) and prothrombin G20210A (RefSeq NM_000506.3, c.*97G>A) mutations, 2-fold for non-O blood group, and 1.5-fold for *FGG* rs2066865 [7].

The genome-wide association strategy is a powerful method to identify common SNPs associated with a complex disorder without a pre-specified hypothesis. Although previous genome-wide association studies (GWASs) have been reported for VTE, none has included more than 1,961 case subjects [5, 6] and none has yielded new genetic loci. In this article, we report the largest investigation to date of the influence of common genetic variations on VTE risk by meta-analyzing GWAS findings from 12 studies.

Subjects and Methods

Study Design

We report on a three-stage investigation of common genetic predictors of VTE. A discovery phase included 7,507 VTE case subjects and 52,632 control subjects

from 12 studies and a replication phase included 3,009 case subjects and 2,586 control subjects from three independent studies. In addition, confirmed discoveries were then examined for association with quantitative biomarkers of VTE risk, gene expression in various tissues and cell types, as well as with whole blood DNA methylation levels.

Participants in Discovery and Replication

For the discovery phase, participants were European-ancestry adults in two French case-control studies, two Dutch case-control studies, and four cohort and four case-control studies from the United States. Details of each study have been previously published [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19]. Three other French case-control studies for VTE were used for the replication stage [11, 20]. In all studies, VTE (PE or DVT) was objectively diagnosed by physicians using different techniques including compression venous duplex ultrasonography, computed tomography, Doppler ultrasound, impedance plethysmography, magnetic resonance, venography, pulmonary angiography, and ventilation/perfusion lung scan. VTE events related to cancer, autoimmune disorders, or natural anticoagulant inhibitor deficiencies (protein C, protein S, antithrombin) were excluded in most studies. A detailed description of the design and the clinical characteristics of all VTE studies analyzed in this work is presented in Table S1. All participating studies were approved by their respective institutional review board and informed consent was obtained from studied individuals.

Genotyping and Imputation

Within each discovery study, DNA samples were genotyped with high-density SNP arrays and were imputed for SNPs available in the 1000 Genomes reference dataset. Summary descriptions of genotyping technologies, quality-control procedures, and imputation methods used for the discovery cohorts are shown in Table S1. In the replication studies, genotyping of the selected SNPs was performed by allele-specific PCR.

Association Analyses and Meta-Analysis for Discovery

Association analyses of imputed SNPs with VTE risk were performed separately in each study by using logistic or Cox-proportional regression analyses adjusted for study-specific covariates (Table S1). All SNPs with acceptable imputation quality ($r^2 > 0.3$) [21] in all 12 discovery studies and with estimated minor allele frequency greater than 0.005 were entered into a meta-analysis. For the meta-analysis, a fixed-effects model based on the inverse-variance weighting was employed as implemented in the METAL software [22]. A statistical threshold of $5 * 10^{-8}$ controlling for the number of independent tests across the genome was applied to declare genome-wide significance [21, 23, 24]. Heterogeneity of the SNP associations across studies was tested with the Cochran's Q statistic and its magnitude expressed by the I^2 index.

Association Analyses and Meta-Analysis for Replication

In the replication cohorts, association of tested SNPs with VTE risk was assessed by use of logistic regression under the assumption of additive allele effects, adjusting

for age and sex. Results obtained in the replication cohorts were meta-analyzed via a fixed-effects model based on the inverse-variance weighting as implemented in the METAL software [22]. A statistical threshold of 0.05 divided by the number of replications performed was used to declare statistical replication. Heterogeneity of the SNP associations across studies and subgroups of individuals (e.g., PE versus DVT) was tested via the Cochran's Q statistic and its magnitude expressed by the I^2 index. For replicated SNPs, meta-analyses of all studies were performed to produce the most robust estimate of the effect size.

Conditional Analysis to Discover Independent Signals at Replicated Loci

To test for the presence of additional independent VTE-associated SNPs at each of the replicated loci, we re-analyzed in each discovery study a GWAS conditioning on the imputed allelic dose of each replicated SNP and meta-analyzed the results by the same strategy as the original GWAS analysis. Areas within 200 kb up- and downstream of the SNPs were examined.

Biologic Follow-up on Replicated Findings

The influence of replicated SNPs on the variability of hemostatic traits known to associate with VTE pathophysiology and measured in our available cohorts was assessed to learn more about biologic pathways. Investigated quantitative biomarkers were D-dimers, endogenous thrombin generation, plasma antigen or activity levels of fibrinogen, coagulation factors II, VII, VIII, IX, X, and XII, von Willebrand factor, antithrombin, protein C, protein S (total and free), protein Z, activated partial thromboplastin time, hemoglobin, and white blood cell and platelet counts. Associations of replicated SNPs with these quantitative hemostatic traits were investigated in five cohorts that were part of the discovery and replication stages by using linear regression analyses adjusted for age-, sex-, and cohort-specific covariates. The overall statistical evidence for association with a given phenotype across available cohorts was assessed by use of the Fisher combined test statistics to account for different measurement methods used across studies.

Replicated SNPs were examined for association with the expression of their structural genes via publicly available genome-wide gene expression data from multiple cell lines and tissues. Impact of replicated lead SNPs on DNA methylation levels from peripheral blood DNA was also investigated.

Results

After applying quality-control measures, 6,751,884 SNPs were tested for association with VTE in a total of 7,507 case subjects and 52,632 control subjects. The Manhattan and Q-Q plots of the meta-analysis of GWAS results are shown in Figures S1–S4. A total of 1,060 SNPs clustered into nine chromosomal regions and reached the genome-wide significance level of $p < 5 * 10^{-8}$.

Among the nine loci that were genome-wide significant in the discovery scan, six were already known to be associated with VTE (*ABO* [MIM 110300], *F2*, *F5*, *F11* [MIM 264900], *FGG*, and *PROCR* [MIM 600646]) whereas three had not been previously reported to be associated with VTE: *TSPAN15* (MIM 613140), *SLC44A2* (MIM 606106), and *ZFPM2* (MIM 603693) (Table 1). At loci with SNPs known to

be associated with VTE, we did not identify additional VTE associations with variants. Detailed descriptions of findings at the known loci are in Figures S1 and S3.

Table 1 Main Findings at Loci Demonstrating Genome-wide Significant Association with VTE in the Discovery Cohorts

				Discovery Stage (7,507 Case Subjects and 52,632 Control Subjects)				Replication Stage (3,009 Case Subjects and 2,586 Control Subjects)			Combined
Lead SNP	Chr.	Gene	Description	Risk Allele	Risk Allele Frequency	Allelic OR [a]	p	Risk Allele Frequency	Allelic OR [a]	p	p [b]
Known Loci											
rs6025	1	*F5*	missense	T	0.033	3.25 (2.91-3.64)	$1.10 * 10^{-96}$	-	-	-	-
rs4524 [c]	1	*F5*	missense	T	0.736	1.20 (1.14-1.26)	$2.65 * 10^{-11}$	-	-	-	-
rs2066865	4	*FGG*	3' UTR	A	0.244	1.24 (1.18-1.32)	$1.03 * 10^{-16}$	-	-	-	-
rs4253417	4	*F11*	intronic	C	0.405	1.27 (1.22-1.34)	$1.21 * 10^{-23}$	-	-	-	-
rs529565	9	*ABO*	intronic	C	0.354	1.55 (1.48-1.63)	$4.23 * 10^{-75}$	-	-	-	-
rs1799963	11	*F2*	intronic	A	0.010	2.29 (1.75-2.99)	$1.73 * 10^{-9}$	-	-	-	-
rs6087685	20	*PROCR*	intronic	C	0.302	1.15 (1.10-1.21)	$1.65 * 10^{-8}$	-	-	-	-
New Loci											
rs4602861	8	*ZFPM2*	intronic	A	0.766	1.20 (1.13-1.27)	$3.48 * 10^{-9}$	0.714	1.02 (0.94-1.11)	0.631	$5.04 * 10^{-7}$
rs78707713	10	*TSPAN15*	intronic	T	0.878	1.28 (1.19-1.39)	$5.74 * 10^{-11}$	0.891	1.42 (1.24-1.62)	$2.21 * 10^{-7}$	$1.67 * 10^{-16}$
rs2288904 [d]	19	*SLC44A2*	intronic	G	0.785	1.19 (1.12-1.26)	$1.07 * 10^{-9}$	0.764	1.28 (1.16-1.40)	$2.64 * 10^{-7}$	$2.75 * 10^{-15}$

[a] Allelic odds ratio associated with the risk allele with its corresponding confidence interval. For ease of presentation, all confidence intervals shown in this table have been computed at a $\alpha = 0.05$.
[b] The combined p value was derived from a fixed effect meta-analysis of the discovery and replication results.
[c] The rs4524 was associated with VTE risk independently of the rs6025 variant.
[d] The rs2288904 was not the lead SNP observed at the *SLC44A2* locus. However, because it is a non-synonymous polymorphism with strong evidence for functionality that is in strong LD ($r^2 \sim 1$) with the lead intronic rs2360742 ($p = 5.6 * 10^{-10}$), it was the one taken forward for replication.

For the three unknown loci, the region-specific lead SNPs were all common intronic variants: rs78707713 in *TSPAN15* with risk allele frequency (RAF) 0.88 and associated odds ratio (OR) for VTE of 1.28 ($p = 5.74 * 10^{-11}$), rs2360742 in *SLC44A2* with RAF = 0.78 and OR = 1.20 ($p = 5.59 * 10^{-10}$), and rs4602861 in *ZFPM2* with RAF = 0.77 and OR = 1.20 ($p = 3.48 * 10^{-9}$). These associations showed little heterogeneity across studies, with Cochran's Q = 16.8, $I^2 = 0.34$, p = 0.11 for *TSPAN15* rs78707713. Corresponding values were Q = 5.64, $I^2 = 0.00$, p = 0.90 for *SLC44A2* rs2360742 and Q = 15.2, $I^2 = 0.28$, p = 0.17 for *ZFPM2* rs4602861. Of note, the intronic *SLC44A2* rs2360742 was in complete association ($r^2 = 1$) with the non-synonymous rs2288904 (p.Arg154Gln [c.461G>A]) that ranked ninth (in terms of p value) at this locus ($p = 1.07 * 10^{-9}$). No coding SNPs in *TSPAN15* or *ZFPM2* were in high linkage disequilibrium (LD) with their lead SNPs.

We attempted replication of *TSPAN15* rs78707713, *SLC44A2* rs2288904, and *ZFPM2* rs4602861 SNPs in the meta-analysis of three independent studies totaling 3,009 VTE-affected individuals and 2,586 control subjects. We did not observe any evidence for association of *ZFPM2* rs4602861 with VTE risk, overall or in any of the three replication studies (Table 2), despite having 93% power in the combined replication studies to detect at the 0.017 (= 0.05/3) statistical threshold the effect of a SNP with RAF 0.76 and associated with an OR of 1.20 [25]. Conversely, we confirmed association for *TSPAN15* rs78707713 and *SLC44A2* rs2288904 SNPs (Table 2). In meta-analyzed replication results, the common *TSPAN15* rs78707713-T allele was associated with an OR for VTE of 1.42 ($p = 2.21 * 10^{-7}$) (Figure S5), and the common *SLC44A2* rs2288904-G allele with an increased risk of 1.28 ($p = 2.64 * 10^{-7}$) (Figure S6). When the results obtained in the discovery and the replication studies were meta-analyzed, the summary OR for VTE was 1.31 ($p = 1.67 * 10^{-16}$) for *TSPAN15* rs78707713 (Figure S5) and 1.21 ($p = 2.75 * 10^{-15}$) for *SLC44A2* rs2288904 (Figure S6). No heterogeneity across the 15 studies was present: Q = 18.6, $I^2 = 0.25$, and p = 0.18 for *TSPAN15* rs78707713 and Q = 7.40, $I^2 = 0.00$, and p = 0.92 for *SLC44A2* rs2288904.

Table 2 Replication Findings for Three SNPs that Reached Genome-wide Significance in the Discovery GWAS

	SLC44A2 rs2288904			*TSPAN15* rs78707713			*ZFPM2* rs4602861		
	GG	GA	AA	TT	TC	CC	AA	AG	GG
MARTHA12									
Control subjects	468 (60%)	258 (33%)	58 (7%)	624 (80%)	146 (19%)	5 (1%)	409 (53%)	314 (40%)	56 (7%)
Case subjects	764 (64%)	391 (33%)	35 (3%)	1,025 (86%)	162 (13%)	7 (¡1%)	629 (53%)	472 (39%)	92 (8%)
RAF [a]	0.762 versus 0.806			0.899 versus 0.926			0.727 versus 0.745		
Allelic OR [b]	1.302 (1.115-1.519); $p = 8.67 * 10^{-4}$			1.428 (1.136-1.798); $p = 2.31 * 10^{-3}$			1.002 (0.866-1.157); $p = 0.986$		
FARIVE									
Control subjects	339 (59%)	204 (35%)	32 (6%)	452 (79%)	114 (20%)	9 (1%)	284 (50%)	231 (40%)	60 (10%)
Case subjects	370 (64%)	189 (32%)	22 (4%)	478 (82%)	102 (18%)	1 (¡1%)	289 (50%)	243 (42%)	48 (8%)
RAF	0.767 versus 0.799			0.885 versus 0.910			0.695 versus 0.708		
Allelic OR	1.208 (0.989-1.475); $p = 0.064$			1.317 (1.000-1.733); $p = 0.050$			1.064 (0.891-1.271); $p = 0.495$		
EDITH									
Control subjects	680 (58%)	422 (36%)	66 (6%)	921 (79%)	236 (20%)	14 (1%)	564 (49%)	487 (42%)	110 (9%)
Case subjects	739 (63%)	394 (34%)	30 (3%)	993 (85%)	172 (15%)	8 (¡1%)	570 (49%)	478 (41%)	110 (10%)
RAF	0.763 versus 0.805			0.887 versus 0.920			0.695 versus 0.699		
Allelic OR	1.294 (1.121-1.492); $p = 4.32 * 10^{-4}$			1.457 (1.197-1.776); $p = 1.76 * 10^{-4}$			1.014 (0.896-1.148); $p = 0.821$		
Combined allelic OR [c]	1.277 (1.164-1.403); $p = 2.64 * 10^{-7}$			1.416 (1.241-1.616); $p = 2.21 * 10^{-7}$			1.021 (0.939-1.110); $p = 0.631$		

[a] Risk allele frequencies in control subjects and in case subjects, respectively.
[b] Allelic odds ratio for disease associated with the risk allele of the studied polymorphisms derived from a logistic regression analysis adjusted for age and sex.
[c] Combined adjusted allelic OR derived from a standard fixed-effect meta-analysis of the results observed in the three replication studies. There was no evidence for heterogeneity across the three replication studies for the association of rs2288904 ($I^2 = 0.388$, p = 0.823), rs78707713 ($I^2 = 0.364$, p = 0.833), or rs4602861 ($I^2 = 0.285$, p = 0.867) with the disease.

Visual inspection of the regional association plots at the *TSPAN15* (Figure 1) and *SLC44A2* (Figure 2) suggested that all SNPs at these loci with high level of significance for association with VTE were in strong LD with the lead SNPs. This was confirmed by the results of the conditional analysis that did not produce any other signal significant at the $5 * 10^{-8}$ statistical threshold. After adjusting for rs78707713, the lowest p value observed at *TSPAN15* was $p = 4.92 * 10^{-4}$ for the intronic rs1072160 (Figure 1). Conditioning on rs2288904 abolished all associations at the *SLC44A2* locus, the lowest p value then being p = 0.047 for the *KEAP1* rs45524632 (Figure 2).

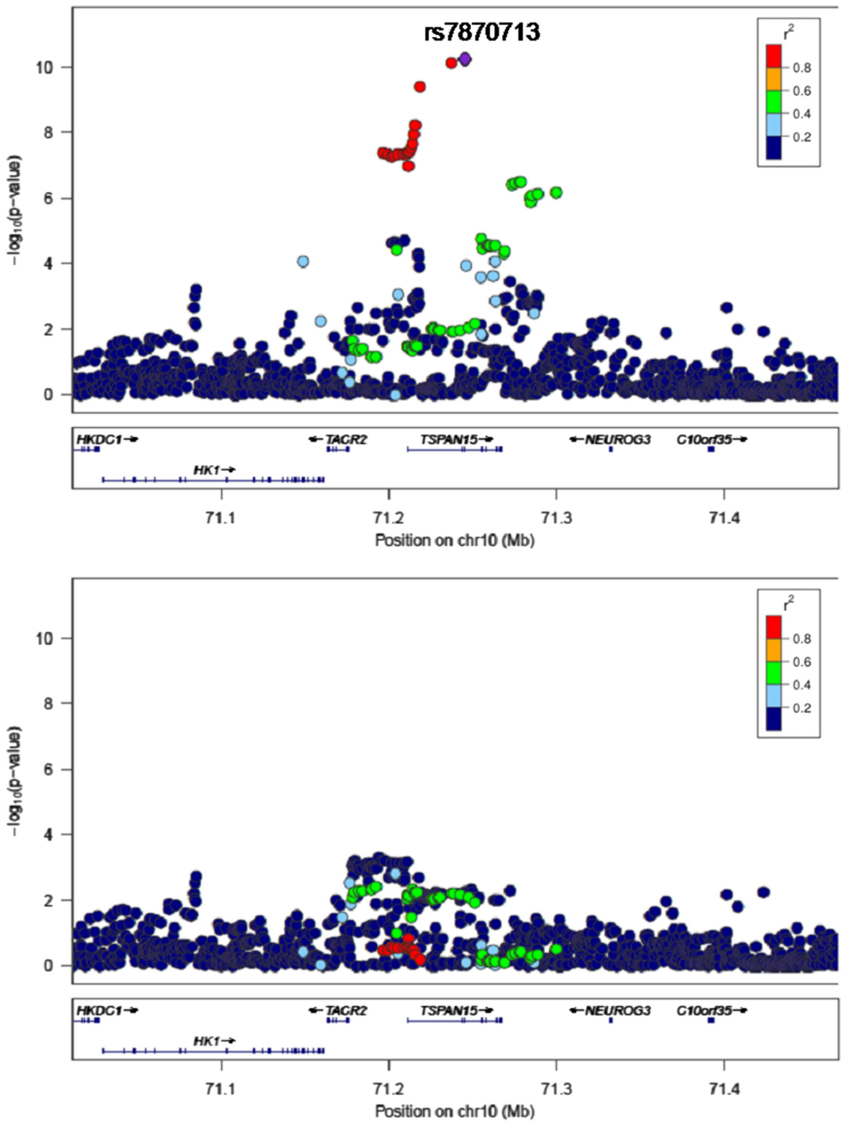

Figure 1 Regional Association Plot at the TSPAN15 Locus.
Association results were derived from the meta-analysis of 12 GWASs (top) that were further conditioned on the effect of *TSPAN15* rs7870713 (bottom). SNPs are colored according to their pairwise LD r^2 with the lead rs7870713. r^2 was estimated from 1000 Genomes (Mar 2012 Eur) database.

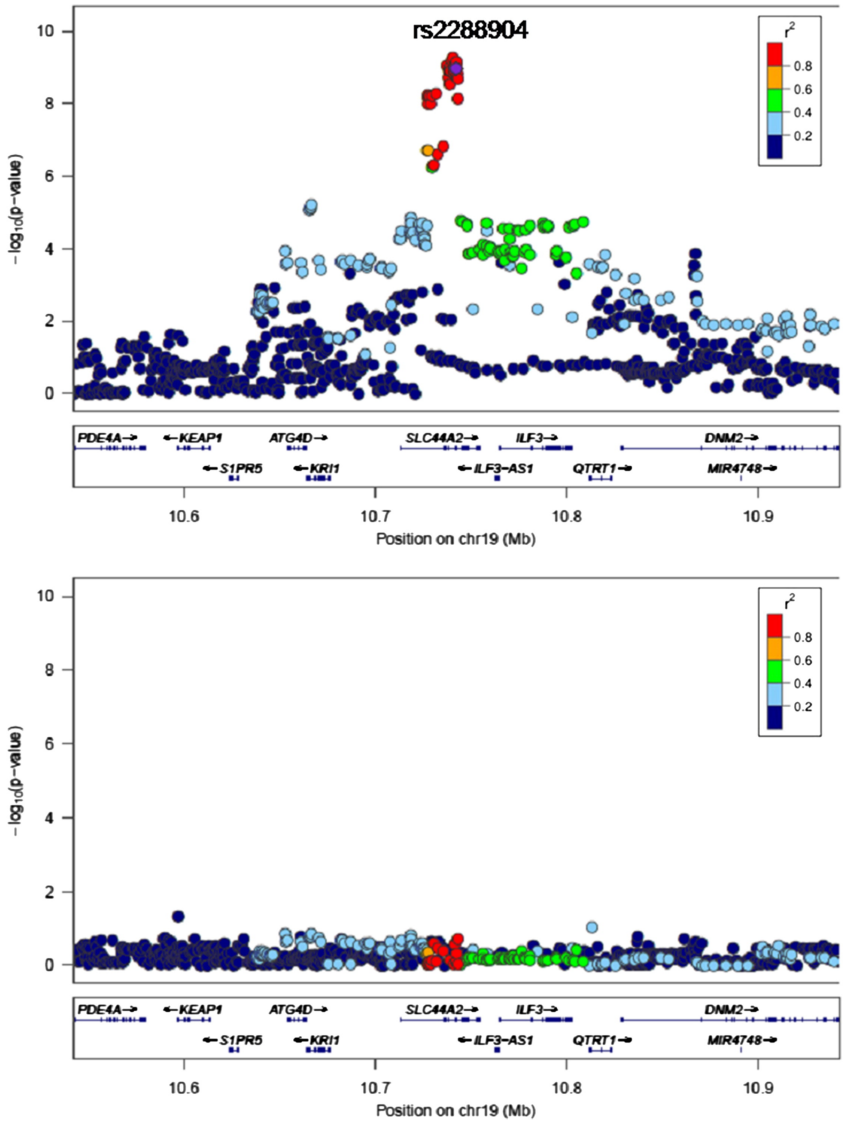

Figure 2 Regional Association Plot at the SLC44A2 Locus.
Association results were derived from the meta-analysis of 12 GWASs (top) that were further conditioned on the effect of *SLC44A2* rs2288904 (bottom). SNPs are colored according to their pairwise LD r^2 with the lead rs2288904. r^2 was estimated from 1000 Genomes (Mar 2012 Eur) database.

In subgroup analyses of the replication studies, the genotype distribution of *TSPAN15* rs78707713 and *SLC44A2* rs2288904 did not differ according to the clinical manifestations of VTE, either PE or DVT (Table S2). We did not observe heterogeneity in the effects of the two SNPs according to sex nor to *F5* rs6025 or *F2* rs1799963 mutations (Tables S3 and S4).

Although the six known VTE-associated loci affect VT risk through a modulation of known hemostatic traits (i.e., levels of von Willebrand factor and Factor VIII for *ABO*, of FXI for *F11*, of endogenous thrombin potential for *F2*, of resistance to activated protein C for *F5*, of fibrinogen for *FGG*, and of protein C for *PROCR*

[26, 27, 28, 29, 30, 31, 32, 33]), the loci of the two discovered genetic associations were not in or near genes that are currently known to influence hemostasis. We then explored whether these SNPs were associated with well-characterized hemostasis phenotypes that are associated with thrombotic propensity. We did not find any evidence for an association with 25 plasma biomarkers (Table S5) even though the sample size of the investigated studies were sufficiently statistically powered ($\sim$ 95%) to detect the additive allele effect of any SNP that would explain 1% of the variability of a quantitative trait [34]. By contrast, the anticipated effects of the six known VTE-associated loci were observed in these studies (Table S6).

We interrogated genome-wide gene expression in ten tissues and DNA methylation studies in blood to determine whether the *TSPAN15* and *SLC44A2* SNPs influenced the regulation of their associated genes. We observed significant associations of the *TSPAN15* rs78707713 with *TSPAN15* DNA methylation measured from peripheral blood DNA and *TSPAN15* expression in macrophages, endothelial cells, and esophagus mucosa (Table 3). The *SLC44A2* rs2288904 was found significantly associated with *SLC44A2* gene expression in monocytes, macrophages, and whole blood (Table 4). However, for none of the interrogated bio-resources for gene expression or DNA methylation levels did the identified disease-associated SNPs show the strongest locus-wide effects. Of note, most SNPs that demonstrated the greatest influence on gene regulation at the replicated loci showed weak or null associations with VTE risk (Tables 3 and 4). Further analyses revealed that the observed effects of *TSPAN15* rs78707713 and *SLC44A2* rs2288904 on gene expression and DNA methylation were probably due to their linkage disequilibrium with other regulatory variants with stronger effects (data not shown).

Table 3 *TSPAN15* SNPs Showing the Strongest Influence on *TSPAN15* Gene Expression and DNA Methylation Levels in Various Human Bio-resources and Their Relation with the *TSPAN15* VTE-Associated rs78707713

			Best *cis* eQTL/mQTL SNP			rs78707713	
Bio-resource	Cell Type or Tissue	Interrogated Bio-resources	rsID	eQTL/mQTL P-value	VTE Association P-value	eQTL/mQTL P-value	r^2 between rs78707713 and Best rsID [a]
Gene expression	B cells	Fairfax et al. [35]	rs7897621	0.0053	0.267	0.129	0.001
	endothelial cells	Erbilgin et al. [36]	rs768498	$6.87 * 10^{-27}$	0.807	$2.80 * 10^{-11}$	0.180
	esophagus mucosa	GTEx Consortium [37]	rs28463525	$2.2 * 10^{-17}$	$5.64 * 10^{-9}$	$2.4 * 10^{-16}$	0.841
	heart, left ventricle	GTEx Consortium [37]	rs10823376	$5.9 * 10^{-9}$	0.911	NA	0.217
	heart	Folkersen et al. [38]	rs768498	$9.25 * 10^{-10}$	0.807	0.038 [b]	0.180
	intestine	Kabakchiev et al. [39]	rs4565792	$2.28 * 10^{-5}$	0.639	0.373	0.242
	nerve-tibial	Folkersen et al. [38]	rs34187097	$5.8 * 10^{-10}$	0.507	NA	0.000
	liver	Folkersen et al. [38]	rs2084274	$9.02 * 10^{-4}$	0.195	0.092 [b]	0.011
	liver	Schadt et al. [40]	rs972570	$3.16 * 10^{-30}$	$3.31 * 10^{-4}$	NA	0.196
	macrophages	Garnier et al. [41]	rs768498	$1.85 * 10^{-39}$	0.807	$1.12 * 10^{-6}$	0.180
	monocytes	Fairfax et al. [35]	rs2812541	0.012	0.356	0.991	0.013
	monocytes	Garnier et al. [41]	rs10128334	$8.05 * 10^{-3}$	0.867	0.978	0.004 *
	2 hr LPS-stimulated monocytes	Fairfax et al. [42]	rs10823371	$3.50 * 10^{-4}$	0.701	0.233	0.217
	24 hr LPS-stimulated monocytes	Fairfax et al. [42]	rs1052179	$2.90 * 10^{-3}$	0.073	0.693	0.148
	2 hr IFN-stimulated monocytes	Fairfax et al. [42]	rs5030949	$1.65 * 10^{-3}$	0.025	0.307	0.004
DNA methylation	whole blood	Dick et al. [43]	rs12416520	$2.24 * 10^{-187}$	0.035	$3.53 * 10^{-46}$	0.489

[a] Pairwise r^2 was derived from the SNAP database [44] except for result noted with asterisk (*) where linkage disequilibrium was estimated from the MARTHA GWAS imputed genotypes.
[b] In the ASAP study, the rs12242391 served as a proxy ($r^2 = 0.84$) for rs78707713 that was not typed.

Table 4 *SLC44A2* SNPs Showing the Strongest Influence on *SLC44A2* Gene Expression in Various Human Bio-resources and Their Relation with the *SLC44A2* VTE-Associated rs2288904

		Best *cis* eQTL SNP			rs2288904	
Cell Type or Tissue	Interrogated Bio-resources	rsID	eQTL P-value	VTE Association P-value	eQTL P-value	r^2 between rs2288904 and Best eSNP [a]
B cells	Fairfax et al. [35]	rs8106664	$3.99 * 10^{-4}$	$6.12 * 10^{-9}$	$2.72 * 10^{-3}$	0.906
Heart	Folkersen et al. [38]	rs3760648	$8.59 * 10^{-4}$	0.293	0.868	0.000
Intestine	Kabakchiev et al. [39]	rs11672431	$5.71 * 10^{-5}$	$9.70 * 10^{-3}$	0.050	0.116
Liver	Schadt et al. [40]	rs7251213	$4.01 * 10^{-6}$	$3.57 * 10^{-3}$	NA	0.160
Liver	Folkersen et al. [38]	rs11672431	$5.29 * 10^{-4}$	$9.70 * 10^{-3}$	0.115	0.116
Macrophages	Garnier et al. [41]	rs3859514	$1.92 * 10^{-12}$	$1.34 * 10^{-4}$	$1.36 * 10^{-9}$	0.393
Monocytes	Garnier et al. [41]	rs62129987	$4.85 * 10^{-25}$	$1.64 * 10^{-5}$	$2.71 * 10^{-12}$	0.440 *
Monocytes	Fairfax et al. [35]	rs7252007	$7.39 * 10^{-30}$	$6.98 * 10^{-3}$	$5.01 * 10^{-9}$	0.345
24 hr LPS-stimulated monocytes	Fairfax et al. [42]	rs7252007	$4.03 * 10^{-10}$		$8.47 * 10^{-5}$	
2 hr LPS-stimulated monocytes	Fairfax et al. [42]	rs7252007	$1.06 * 10^{-5}$		$4.70 * 10^{-4}$	
2 hr IFN-stimulated monocytes	Fairfax et al. [42]	rs12609501	$1.92 * 10^{-12}$	$9.23 * 10^{-3}$	$1.57 * 10^{-7}$	0.072
Whole blood	Westra et al. [45]	rs892078	$3.64 * 10^{-93}$	$2.11 * 10^{-5}$	$3.85 * 10^{-70}$	0.438

Of note, no CpG probe targeting the *SLC44A2* locus satisfied the adopted quality-control procedures, preventing us from efficiently testing for the effect of *SLC44A2* rs2288904 on DNA methylation levels at this locus in the interrogated bio-resources.

[a] Pairwise r^2 was derived from the SNAP database [44] except for result noted with asterisk (*) where linkage disequilibrium was estimated from the MARTHA GWAS imputed genotypes.

Discussion

By using data from more than 7,000 case subjects and more than 50,000 control subjects, which represents a 4-fold increase in the number of VTE events used in our discovery effort compared with that used in the latest meta-analysis of GWASs for VTE, we identified and then replicated two associations for common variants in *TSPAN15* and *SLC44A2*. The strengths of the associations were modest in size with ORs of 1.3 or less, but the statistical evidence was robust in the discovery and replication stages. Further, we observed that the variants were not associated with dozens of hemostatic markers characterizing the coagulation/fibrinolysis balance. The identified VTE-associated SNPs map to genes that are not in conventional pathways to thrombosis that have marked most of the genetic associations to date, suggesting that these genetic variants represent novel biological pathways leading to VTE.

The common T allele, with frequency $\sim$ 0.89, of the identified *TSPAN15* rs78707713 was associated with an increased risk of 1.31-fold. The *TSPAN15* rs78707713 is intronic and the interrogation of several gene expression databases as well as the application of prediction/annotation tools [46, 47, 48] did not suggest any regulatory elements supporting a functional role of this SNP. It is likely that the SNP is in strong LD with yet unidentified culprit variant(s). For instance, rs78707713 is in strong LD (pairwise $r^2 = 0.89$) with the intronic *TSPAN15* rs17490626 predicted [46, 47] to map an enhancer domain, the significance of the VTE association of the latter SNP being $p = 3.74 * 10^{-10}$. No association of *TSPAN15*-coding SNPs with VTE risk was observed in the discovery meta-analysis (smallest p value = 0.49). *TSPAN15* codes for tetraspanin 15, a member of the tetraspanin superfamily that act as scaffolding proteins, anchoring multiple proteins to the cell membrane [49]. Members of the tetraspanin family have roles in cells that regulate hemostasis. *TSPAN24* (CD151)- [50] and *TSPAN32* (TSSC6)- [51] deficient mice exhibit a bleeding phenotype with impaired "outside-in" signaling through αIIbβ3, the major platelet integrin. CD63 (*TSPAN30*) facilitates the release of von Willebrand factor (*VWF*) from endothelial cell Weibel-Palade bodies through transient enhancement of fusion between the Weibel-Palade body membrane and the plasma membrane [52], which probably contributes to the leucocyte attachment to the endothelium [53].

The risk allele at the second identified locus, *SLC44A2* rs2288904-G, was also common, with frequency $\sim$ 0.77. It was associated with a relative risk of 1.21 and is probably the functional variant. Indeed, the observed risk allele (G) codes for the Arg154 isoform of the choline transporter-like protein 2 (CTL-2). CTL-2 has been associated with several human diseases [54], including transfusion-related acute lung injury (TRALI). TRALI is a life-threatening complication of blood transfusion and the leading cause of transfusion-associated mortality in developed countries. Severe TRALI is due to antibodies in blood components directed against the human neutrophil alloantigen-3a (HNA-3a), which is determined by the Arg154 isoform [55, 56]. Greinacher et al. [55] found that alloantibodies targeting CTL-2 lead to leucocyte activation and aggregation. Recently, human anti-HNA-3a antibodies were shown to directly interact with endothelial CTL-2 to disturb pulmonary endothelial barrier function that would lead to severe TRALI [57]. It might be that

carrying the HNA-3a antigen, determined by the Arg154 isoform, favors activation of leucocytes/neutrophils and endothelial cells in some triggering circumstances.

We did not observe any difference in *TSPAN15* and *SLC44A2* VTE-associated SNP allele frequencies between DVT- and PE-affected individuals in the replication populations, the two main clinical manifestations of VTE. This observation suggests that the underlying pathophysiological mechanisms are more likely to be involved in thrombus formation rather than its rupture and its migration toward the pulmonary vein. About 20% of persons with unprovoked VTE (i.e., occurring without clear external factors like surgery, trauma, immobilization, hormone use, or cancer) will experience a recurrent event, even after a 6-month course of anticoagulant prophylaxis. Because we had little information about recurrence follow-up in our studies, we were not able to assess whether the identified SNPs can discriminate between individuals that will and those that will not face a recurrent event. Investigating whether these SNPs can help improve the secondary prevention of VTE would definitively be warranted.

Despite having gathered the largest GWAS samples of VTE-affected individuals, our approach was not well powered to identify common SNPs associated with more modest effects than those observed for *SLC44A2* and *TSPAN15*, i.e., with OR $<$ 1.20, or extremely rare mutations (e.g., with frequency $< 1\%$) that cannot be efficiently tagged by the tested SNPs (Table S7). Heterogeneity in the design and clinical characteristics of the studied populations might also have contributed to slightly attenuate the power of our study to detect additional genome-wide significant associations. Further investigations deserve to be conducted to better characterize the 202 suggestive statistical associations with $p < 10^{-5}$ in our discovery meta-analysis and to identify additional susceptibility loci for VTE.

Though additional work is needed, including the identification of the functional *TSPAN15* variant(s), our results demonstrated that *SLC44A2* and *TSPAN15* are two susceptibility loci for VTE. The products of the two genes, expressed by cells that are central to the pathophysiology of thrombosis (monocytes/macrophages and endothelial cells), are not known to have central roles in the traditional hemostasis pathways nor to other cardiovascular diseases. These results pave the way for novel mechanistic concepts of VTE pathophysiology, new biomarkers for the disease, and novel therapeutic perspectives.

Supplemental Data

Supplemental Data include six figures, ten tables, Supplemental Acknowledgments, funding information for each cohort, and a list of members of the INVENT consortium and can be found with this article online at http://dx.doi.org/10.1016/j.ajhg.2015.01.019.

Acknowledgements

This work is the product of the International Network against Thrombosis (INVENT) collaboration. This work was financially supported by the French National Institute for Health and Medical Research; French Ministry of Health; French Medical Research Foundation; French Agency for Research; National Heart, Lung, and Blood Institute (USA); National Institute for Health Research (USA); National Human Genome Research Institute (USA); National Cancer Institute (USA); the Donald W. Reynolds Foundation; Fondation Leducq; The European Commission; Netherlands Organisation for Scientific Research; Netherlands Heart Foundation; Dutch Cancer Foundation; Canadian Institutes of Health Research; Heart and Stroke Foundation of Canada; British Heart Foundation; and ICAN Institute for Cardiometabolism and Nutrition. The lists of grants supporting this research can be found in the Supplemental Data.

Web Resources

The URLs for data presented herein are as follows:
1000 Genomes, http://browser.1000genomes.org
OMIM, http://www.omim.org/
RefSeq, http://www.ncbi.nlm.nih.gov/RefSeq

Author details

[1] Institut National pour la Santé et la Recherche Médicale (INSERM), Unité Mixte de Recherche en Santé (UMR_S) 1166, 75013 Paris, France. [2] Sorbonne Universités, Université Pierre et Marie Curie (UPMC Univ Paris 06), UMR_S 1166, Team Genomics & Pathophysiology of Cardiovascular Diseases, 75013 Paris, France. [3] Institute for Cardiometabolism and Nutrition (ICAN), 75013 Paris, France. [4] Division of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, MA 02215 Boston, USA. [5] Department of Thrombosis and Hemostasis, Department of Clinical Epidemiology, Leiden University Medical Center, 2333 ZA Leiden, the Netherlands. [6] Division of Epidemiology and Community Health, University of Minnesota, MN 55454 Minneapolis, USA. [7] Program in Genetic Epidemiology and Statistical Genetics, Department of Epidemiology, Harvard School of Public Health, MA 02115 Boston, USA. [8] Division of Biomedical Statistics and Informatics, Mayo Clinic, MN 55905 Rochester, USA. [9] Einthoven Laboratory for Experimental Vascular Medicine, Department of Thrombosis and Hemostasis, Leiden University Medical Center, 2300 RC Leiden, the Netherlands. [10] Department of Epidemiology, University of Washington, WA 98195 Seattle, USA. [11] Laboratory of Haematology, La Timone Hospital, 13385 Marseille, France. [12] INSERM, UMR_S 1062, Nutrition Obesity and Risk of Thrombosis, 13385 Marseille, France. [13] Nutrition Obesity and Risk of Thrombosis, Aix-Marseille University, UMR_S 1062, 13385 Marseille, France. [14] Université Paris Descartes, Sorbonne Paris Cité, 75006 Paris, France. [15] AP-HP, Hopital Européen Georges Pompidou, Service d'Hématologie Biologique, 75015 Paris, France. [16] Faculté de Pharmacie, INSERM, UMR_S 1140, 75006 Paris, France. [17] Université de Brest, EA3878 and CIC1412, 29238 Brest, France. [18] Ottawa Hospital Research Institute at the University of Ottawa, ON K1Y 4E9 Ottawa, Canada. [19] Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, NY 10029 New York, USA. [20] Department of Cardiovascular Sciences, University of Leicester, LE1 7RH Leicester, UK. [21] National Institute for Health Research (NIHR) Leicester Cardiovascular Biomedical Research Unit, LE3 9QP Leicester, UK. [22] Department of PharmacoGenetics, Novo Nordisk Park 9.1.21, 2400 Copenhagen, Denmark. [23] Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, 2300 RC Leiden, the Netherlands. [24] Division of Preventive Medicine, Brigham and Women's Hospital, MA 02215 Boston, USA. [25] Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, WA 98195-5852 Seattle, USA. [26] Department of Molecular Epidemiology, Leiden University Medical Center, 2300 RC Leiden, the Netherlands. [27] Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, ON M5T 3M7 Toronto, Canada. [28] Commissariat à l'Energie Atomique/Direction des Sciences du Vivant/Institut de Génomique, Centre National de Génotypage, 91057 Evry, France. [29] William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, E1 4NS London, UK. [30] Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, 21589 Jeddah, Saudi Arabia. [31] Inserm Research Center U897, University of Bordeaux, 33000 Bordeaux, France. [32] Inserm Research Unit U1061, University of Montpellier I, 34000 Montpellier, France. [33] Los Angeles Biomedical Research Institute and Department of Pediatrics, Harbor-UCLA Medical Center, CA 90502 Torrence, USA. [34] Department of Medicine, University of California, CA 90095 Los Angeles, USA. [35] Atherosclerosis Research Unit, Center for Molecular Medicine, Department of Medicine, Karolinska Institutet, 171 77 Stockholm, Sweden. [36] Group Health Research Institute, Group Health Cooperative, WA 98101 Seattle, USA. [37] Institut Pasteur de Lille, Université de Lille Nord de France, INSERM UMR_S 744, 59000 Lille, France. [38] Centre Hospitalier Régional Universitaire de Lille, 59000 Lille, France. [39] Division of Biostatistics, University of Minnesota, MN 55455 Minneapolis, USA. [40] Department of Emergency Medicine, Massachusetts General Hospital, Channing Network Medicine, Harvard Medical School, MA 2114 Minneapolis, Boston. [41] Division of Cardiovascular Diseases, Mayo Clinic, MN 55905 Rochester, USA. [42] Seattle Epidemiologic Research and Information Center, VA Office of Research and Development, WA 98108 Rochester, Seattle.

References

1. Næss, I.A., Christiansen, S.C., Romundstad, P., Cannegieter, S.C., Rosendaal, F.R., Hammerstrøm, J.: Incidence and mortality of venous thrombosis: a population-based study **5**(4), 692–699. doi:10.1111/j.1538-7836.2007.02450.x. Accessed 2015-10-19
2. Virchow, R.: Gesammelte Abhandlungen zur Wissenschaftlichen Medicin. Hamm : Grote. http://archive.org/details/gesammelteabhand00virc Accessed 2015-10-19
3. Heit, J.A., Phelps, M.A., Ward, S.A., Slusser, J.P., Petterson, T.M., De Andrade, M.: Familial segregation of venous thromboembolism **2**(5), 731–736. doi:10.1111/j.1538-7933.2004.00660.x. Accessed 2015-10-19
4. Zöller, B., Li, X., Sundquist, J., Sundquist, K.: Age- and gender-specific familial risks for venous thromboembolism a nationwide epidemiological study based on hospitalizations in sweden **124**(9), 1012–1020. doi:10.1161/CIRCULATIONAHA.110.965020. Accessed 2015-10-19
5. Tang, W., Teichert, M., Chasman, D.I., Heit, J.A., Morange, P.-E., Li, G., Pankratz, N., Leebeek, F.W., Paré, G., de Andrade, M., Tzourio, C., Psaty, B.M., Basu, S., Ruiter, R., Rose, L., Armasu, S.M., Lumley, T., Heckbert, S.R., Uitterlinden, A.G., Lathrop, M., Rice, K.M., Cushman, M., Hofman, A., Lambert, J.-C., Glazer, N.L., Pankow, J.S., Witteman, J.C., Amouyel, P., Bis, J.C., Bovill, E.G., Kong, X., Tracy, R.P., Boerwinkle, E., Rotter, J.I., Trégouët, D.-A., Loth, D.W., Stricker, B.H.C., Ridker, P.M., Folsom, A.R., Smith, N.L.: A genome-wide association study for venous thromboembolism: The extended cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium **37**(5), 512–521. doi:10.1002/gepi.21731. Accessed 2015-10-19
6. Germain, M., Saut, N., Oudot-Mellakh, T., Letenneur, L., Dupuy, A.-M., Bertrand, M., Alessi, M.-C., Lambert, J.-C., Zelenika, D., Emmerich, J., Tiret, L., Cambien, F., Lathrop, M., Amouyel, P., Morange, P.-E., Trégouët, D.-A.: Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: A case study on venous thrombosis **7**(6), 38538. doi:10.1371/journal.pone.0038538. Accessed 2015-10-19
7. Morange, P.-E., Trégouët, D.-A.: Current knowledge on the genetics of incident venous thrombosis **11**, 111–121. doi:10.1111/jth.12233. Accessed 2015-10-19

8. Investigators, T.A.: The atherosclerosis risk in communities (ARIC) study: design and objectives. the ARIC investigators **129**(4), 687–702
9. Fried, L.P., Borhani, N.O., Enright, P., Furberg, C.D., Gardin, J.M., Kronmal, R.A., Kuller, L.H., Manolio, T.A., Mittelmark, M.B., Newman, A., O'Leary, D.H., Psaty, B., Rautaharju, P., Tracy, R.P., Weiler, P.G.: The cardiovascular health study: Design and rationale **1**(3), 263–276. doi:10.1016/1047-2797(91)90005-W. Accessed 2015-10-19
10. Tell, G.S., Fried, L.P., Hermanson, B., Manolio, T.A., Newman, A.B., Borhani, N.O.: Recruitment of adults 65 years and older as participants in the cardiovascular health study **3**(4), 358–366. doi:10.1016/1047-2797(93)90062-9. Accessed 2015-10-19
11. Trégouët, D.-A., Heath, S., Saut, N., Biron-Andreani, C., Schved, J.-F., Pernod, G., Galan, P., Drouet, L., Zelenika, D., Juhan-Vague, I., Alessi, M.-C., Tiret, L., Lathrop, M., Emmerich, J., Morange, P.-E.: Common susceptibility alleles are unlikely to contribute as strongly as the FV and ABO loci to VTE risk: results from a GWAS approach **113**(21), 5298–5303. doi:10.1182/blood-2008-11-190389. Accessed 2015-10-19
12. de Visser, M.c.h., van Minkelen, R., van Marion, V., den Heijer, M., Eikenboom, J., Vos, H.l., Slagboom, P.e., Houwing-Duistermaat, J.j., Rosendaal, F.r., Bertina, R.m.: Genome-wide linkage scan in affected sibling pairs identifies novel susceptibility region for venous thromboembolism: Genetics in familial thrombosis study **11**(8), 1474–1484. doi:10.1111/jth.12313. Accessed 2015-10-19
13. Psaty BM, Heckbert SR, Koepsell TD, et al: THe risk of myocardial infarction associated with antihypertensive drug therapies **274**(8), 620–625. doi:10.1001/jama.1995.03530080036038. Accessed 2015-10-19
14. Germain, M., Saut, N., Greliche, N., Dina, C., Lambert, J.-C., Perret, C., Cohen, W., Oudot-Mellakh, T., Antoni, G., Alessi, M.-C., Zelenika, D., Cambien, F., Tiret, L., Bertrand, M., Dupuy, A.-M., Letenneur, L., Lathrop, M., Emmerich, J., Amouyel, P., Trégouët, D.-A., Morange, P.-E.: Genetics of venous thrombosis: Insights from a new genome wide association study **6**(9), 25581. doi:10.1371/journal.pone.0025581. Accessed 2015-10-19
15. Heit, J.A., Armasu, S.M., Asmann, Y.W., Cunningham, J.M., Matsumoto, M.E., Petterson, T.M., De Andrade, M.: A genome-wide association study of venous thromboembolism identifies risk variants in chromosomes 1q24.2 and 9q **10**(8), 1521–1531. doi:10.1111/j.1538-7836.2012.04810.x. Accessed 2015-10-19
16. Blom JW, Doggen CM, Osanto S, Rosendaal FR: MAlignancies, prothrombotic mutations, and the risk of venous thrombosis **293**(6), 715–722. doi:10.1001/jama.293.6.715. Accessed 2015-10-19
17. Hankinson, S.E., Colditz, G.A., Hunter, D.J., Manson, J.E., Willett, W.C., Stampfer, M.J., Longcope, C., Speizer, F.E.: Reproductive factors and family history of breast cancer in relation to plasma estrogen and prolactin levels in postmenopausal women in the nurses' health study (united states) **6**(3), 217–224. doi:10.1007/BF00051793. Accessed 2015-10-19
18. Tworoger, S.S., Sluss, P., Hankinson, S.E.: Association between plasma prolactin concentrations and risk of breast cancer among predominately premenopausal women **66**(4), 2476–2482. doi:10.1158/0008-5472.CAN-05-3369. Accessed 2015-10-19
19. Ridker, P.M., Chasman, D.I., Zee, R.Y.L., Parker, A., Rose, L., Cook, N.R., Buring, J.E., Group, f.t.W.G.H.S.W.: Rationale, design, and methodology of the women's genome health study: A genome-wide association study of more than 25 000 initially healthy american women **54**(2), 249–255. doi:10.1373/clinchem.2007.099366. Accessed 2015-10-19
20. Oger, E., Lacut, K., Le Gal, G., Couturaud, F., Guénet, D., Abalain, J.-H., Roguedas, A.-M., Mottier, D., The Edith Collaborative Study Group: Hyperhomocysteinemia and low b vitamin levels are independently associated with venous thromboembolism: results from the EDITH study: a hospital-based case–control study **4**(4), 793–799. doi:10.1111/j.1538-7836.2006.01856.x. Accessed 2015-10-19
21. Johnson, E.O., Hancock, D.B., Levy, J.L., Gaddis, N.C., Saccone, N.L., Bierut, L.J., Page, G.P.: Imputation across genotyping arrays for genome-wide association studies: assessment of bias and a correction strategy **132**(5), 509–522. doi:10.1007/s00439-013-1266-7. Accessed 2015-10-19
22. Willer, C.J., Li, Y., Abecasis, G.R.: METAL: fast and efficient meta-analysis of genomewide association scans **26**(17), 2190–2191. doi:10.1093/bioinformatics/btq340. Accessed 2015-10-19
23. Li, M.-X., Yeung, J.M.Y., Cherny, S.S., Sham, P.C.: Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets **131**(5), 747–756. doi:10.1007/s00439-011-1118-2. Accessed 2015-10-19
24. Panagiotou, O.A., Ioannidis, J.P.A., Project, f.t.G.-W.S.: What should the genome-wide significance threshold be? empirical replication of borderline genetic associations **41**(1), 273–286. doi:10.1093/ije/dyr178. Accessed 2015-10-19
25. Skol, A.D., Scott, L.J., Abecasis, G.R., Boehnke, M.: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies **38**(2), 209–213. doi:10.1038/ng1706. Accessed 2015-10-19
26. Smith, N.L., Chen, M.-H., Dehghan, A., Strachan, D.P., Basu, S., Soranzo, N., Hayward, C., Rudan, I., Sabater-Lleal, M., Bis, J.C., Maat, M.P.M.d., Rumley, A., Kong, X., Yang, Q., Williams, F.M.K., Vitart, V., Campbell, H., Mälarstig, A., Wiggins, K.L., Duijn, C.M.V., McArdle, W.L., Pankow, J.S., Johnson, A.D., Silveira, A., McKnight, B., Uitterlinden, A.G., Consortium, W.T.C.C., Aleksic, N., Meigs, J.B., Peters, A., Koenig, W., Cushman, M., Kathiresan, S., Rotter, J.I., Bovill, E.G., Hofman, A., Boerwinkle, E., Tofler, G.H., Peden, J.F., Psaty, B.M., Leebeek, F., Folsom, A.R., Larson, M.G., Spector, T.D., Wright, A.F., Wilson, J.F., Hamsten, A., Lumley, T., Witteman, J.C.M., Tang, W., O'Donnell, C.J.: Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von willebrand factor the CHARGE (cohorts for heart and aging research in genome epidemiology) consortium **121**(12), 1382–1392. doi:10.1161/CIRCULATIONAHA.109.869156. Accessed 2015-10-19
27. Li, Y., Bezemer, I.D., Rowland, C.M., Tong, C.H., Arellano, A.R., Catanese, J.J., Devlin, J.J., Reitsma, P.H., Bare, L.A., Rosendaal, F.R.: Genetic variants associated with deep vein thrombosis: the f11 locus **7**(11), 1802–1808. doi:10.1111/j.1538-7836.2009.03544.x. Accessed 2015-10-19
28. Rocanin-Arjo, A., Cohen, W., Carcaillon, L., Frère, C., Saut, N., Letenneur, L., Alhenc-Gelas, M., Dupuy,

A.-M., Bertrand, M., Alessi, M.-C., Germain, M., Wild, P.S., Zeller, T., Cambien, F., Goodall, A.H., Amouyel, P., Scarabin, P.-Y., Trégouët, D.-A., Morange, P.-E., Consortium, a.t.C.: A meta-analysis of genome-wide association studies identifies ORM1 as a novel gene controlling thrombin generation potential **123**(5), 777–785. doi:10.1182/blood-2013-10-529628. Accessed 2015-10-19

29. Oudot-Mellakh, T., Cohen, W., Germain, M., Saut, N., Kallel, C., Zelenika, D., Lathrop, M., Trégouët, D.-A., Morange, P.-E.: Genome wide association study for plasma levels of natural anticoagulant inhibitors and protein c anticoagulant pathway: the MARTHA project **157**(2), 230–239. doi:10.1111/j.1365-2141.2011.09025.x. Accessed 2015-10-19
30. Sabater-Lleal, M., Huang, J., Chasman, D., Naitza, S., Dehghan, A., Johnson, A.D., Teumer, A., Reiner, A.P., Folkersen, L., Basu, S., Rudnicka, A.R., Trompet, S., Mälarstig, A., Baumert, J., Bis, J.C., Guo, X., Hottenga, J.J., Shin, S.-Y., Lopez, L.M., Lahti, J., Tanaka, T., Yanek, L.R., Oudot-Mellakh, T., Wilson, J.F., Navarro, P., Huffman, J.E., Zemunik, T., Redline, S., Mehra, R., Pulanic, D., Rudan, I., Wright, A.F., Kolcic, I., Polasek, O., Wild, S.H., Campbell, H., Curb, J.D., Wallace, R., Liu, S., Eaton, C.B., Becker, D.M., Becker, L.C., Bandinelli, S., Räikkönen, K., Widen, E., Palotie, A., Fornage, M., Green, D., Gross, M., Davies, G., Harris, S.E., Liewald, D.C., Starr, J.M., Williams, F.M.K., Grant, P.J., Spector, T.D., Strawbridge, R.J., Silveira, A., Sennblad, B., Rivadeneira, F., Uitterlinden, A.G., Franco, O.H., Hofman, A., Dongen, J.v., Willemsen, G., Boomsma, D.I., Yao, J., Jenny, N.S., Haritunians, T., McKnight, B., Lumley, T., Taylor, K.D., Rotter, J.I., Psaty, B.M., Peters, A., Gieger, C., Illig, T., Grotevendt, A., Homuth, G., Völzke, H., Kocher, T., Goel, A., Franzosi, M.G., Seedorf, U., Clarke, R., Steri, M., Tarasov, K.V., Sanna, S., Schlessinger, D., Stott, D.J., Sattar, N., Buckley, B.M., Rumley, A., Lowe, G.D., McArdle, W.L., Chen, M.-H., Tofler, G.H., Song, J., Boerwinkle, E., Folsom, A.R., Rose, L.M., Franco-Cereceda, A., Teichert, M., Ikram, M.A., Mosley, T.H., Bevan, S., Dichgans, M., Rothwell, P.M., Sudlow, C.L.M., Hopewell, J.C., Chambers, J.C., Saleheen, D., Kooner, J.S., Danesh, J., Nelson, C.P., Erdmann, J., Reilly, M.P., Kathiresan, S., Schunkert, H., Morange, P.-E., Ferrucci, L., Eriksson, J.G., Jacobs, D., Deary, I.J., Soranzo, N., Witteman, J.C.M., Geus, E.J.C.d., Tracy, R.P., Hayward, C., Koenig, W., Cucca, F., Jukema, J.W., Eriksson, P., Seshadri, S., Markus, H.S., Watkins, H., Samani, N.J., Consortium, V.T.E., Consortium, S., (wtccc2), W.T.C.C.C.., Consortium, C., Consortium, C., Wallaschofski, H., Smith, N.L., Tregouet, D., Ridker, P.M., Tang, W., Strachan, D.P., Hamsten, A., O'Donnell, C.J.: Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease **128**(12), 1310–1324. doi:10.1161/CIRCULATIONAHA.113.002251. Accessed 2015-10-19
31. Smith, N.L., Huffman, J.E., Strachan, D.P., Huang, J., Dehghan, A., Trompet, S., Lopez, L.M., Shin, S.-Y., Baumert, J., Vitart, V., Bis, J.C., Wild, S.H., Rumley, A., Yang, Q., Uitterlinden, A.G., Stott, D.J., Davies, G., Carter, A.M., Thorand, B., Polašek, O., McKnight, B., Campbell, H., Rudnicka, A.R., Chen, M.-H., Buckley, B.M., Harris, S.E., Peters, A., Pulanic, D., Lumley, T., Craen, A.J.M.d., Liewald, D.C., Gieger, C., Campbell, S., Ford, I., Gow, A.J., Luciano, M., Porteous, D.J., Guo, X., Sattar, N., Tenesa, A., Cushman, M., Slagboom, P.E., Visscher, P.M., Spector, T.D., Illig, T., Rudan, I., Bovill, E.G., Wright, A.F., McArdle, W.L., Tofler, G., Hofman, A., Westendorp, R.G.J., Starr, J.M., Grant, P.J., Karakas, M., Hastie, N.D., Psaty, B.M., Wilson, J.F., Lowe, G.D.O., O'Donnell, C.J., Witteman, J.C.M., Jukema, J.W., Deary, I.J., Soranzo, N., Koenig, W., Hayward, C.: Genetic predictors of fibrin d-dimer levels in healthy adults **123**(17), 1864–1872. doi:10.1161/CIRCULATIONAHA.110.009480. Accessed 2015-10-19
32. Tang, W., Schwienbacher, C., Lopez, L., Ben-Shlomo, Y., Oudot-Mellakh, T., Johnson, A., Samani, N., Basu, S., Gögele, M., Davies, G., Lowe, G.O., Tregouet, D.-A., Tan, A., Pankow, J., Tenesa, A., Levy, D., Volpato, C., Rumley, A., Gow, A., Minelli, C., Yarnell, J.G., Porteous, D., Starr, J., Gallacher, J., Boerwinkle, E., Visscher, P., Pramstaller, P., Cushman, M., Emilsson, V., Plump, A., Matijevic, N., Morange, P.-E., Deary, I., Hicks, A., Folsom, A.: Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease **91**(1), 152–162. doi:10.1016/j.ajhg.2012.05.009. Accessed 2015-10-19
33. Tang, W., Basu, S., Kong, X., Pankow, J.S., Aleksic, N., Tan, A., Cushman, M., Boerwinkle, E., Folsom, A.R.: Genome-wide association study identifies novel loci for plasma levels of protein c: the ARIC study **116**(23), 5032–5036. doi:10.1182/blood-2010-05-283739. Accessed 2015-10-19
34. Gauderman, W.J.: Sample size requirements for matched case-control studies of gene–environment interaction **21**(1), 35–50. doi:10.1002/sim.973. Accessed 2015-10-19
35. Fairfax, B.P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F.O., Knight, J.C.: Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles **44**(5), 502–510. doi:10.1038/ng.2205. Accessed 2015-10-19
36. Erbilgin, A., Civelek, M., Romanoski, C.E., Pan, C., Hagopian, R., Berliner, J.A., Lusis, A.J.: Identification of CAD candidate genes in GWAS loci and their expression in vascular cells **54**(7), 1894–1905. doi:10.1194/jlr.M037085. Accessed 2015-10-19
37. Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., Foster, B., Moser, M., Karasik, E., Gillard, B., Ramsey, K., Sullivan, S., Bridge, J., Magazine, H., Syron, J., Fleming, J., Siminoff, L., Traino, H., Mosavel, M., Barker, L., Jewell, S., Rohrer, D., Maxim, D., Filkins, D., Harbach, P., Cortadillo, E., Berghuis, B., Turner, L., Hudson, E., Feenstra, K., Sobin, L., Robb, J., Branton, P., Korzeniewski, G., Shive, C., Tabor, D., Qi, L., Groch, K., Nampally, S., Buia, S., Zimmerman, A., Smith, A., Burges, R., Robinson, K., Valentino, K., Bradbury, D., Cosentino, M., Diaz-Mayoral, N., Kennedy, M., Engel, T., Williams, P., Erickson, K., Ardlie, K., Winckler, W., Getz, G., DeLuca, D., MacArthur, D., Kellis, M., Thomson, A., Young, T., Gelfand, E., Donovan, M., Meng, Y., Grant, G., Mash, D., Marcus, Y., Basile, M., Liu, J., Zhu, J., Tu, Z., Cox, N.J., Nicolae, D.L., Gamazon, E.R., Im, H.K., Konkashbaev, A., Pritchard, J., Stevens, M., Flutre, T., Wen, X., Dermitzakis, E.T., Lappalainen, T., Guigo, R., Monlong, J., Sammeth, M., Koller, D., Battle, A., Mostafavi, S., McCarthy, M., Rivas, M., Maller, J., Rusyn, I., Nobel, A., Wright, F., Shabalin, A., Feolo, M., Sharopova, N., Sturcke, A., Paschal, J., Anderson, J.M., Wilder, E.L., Derr, L.K., Green, E.D., Struewing, J.P., Temple, G., Volpi, S., Boyer, J.T., Thomson, E.J., Guyer, M.S., Ng, C., Abdallah, A., Colantuoni, D., Insel, T.R., Koester, S.E., Little, A.R., Bender, P.K., Lehner, T., Yao, Y., Compton, C.C.,

Vaught, J.B., Sawyer, S., Lockhart, N.C., Demchok, J., Moore, H.F.: The genotype-tissue expression (GTEx) project **45**(6), 580–585. doi:10.1038/ng.2653. Accessed 2015-10-19
38. Folkersen, L., Hooft, F.v., Chernogubova, E., Agardh, H.E., Hansson, G.K., Hedin, U., Liska, J., Syvänen, A.-C., Paulsson-Berne, G., Franco-Cereceda, A., Hamsten, A., Gabrielsen, A., Eriksson, P., Groups, o.b.o.t.B.a.A.s.: Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease **3**(4), 365–373. doi:10.1161/CIRCGENETICS.110.948935. Accessed 2015-10-19
39. Kabakchiev, B., Silverberg, M.S.: Expression quantitative trait loci analysis identifies associations between genotype and gene expression in human intestine **144**(7), 1488–14963. doi:10.1053/j.gastro.2013.03.001. Accessed 2015-10-19
40. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J.D., Avila-Campillo, I., Kruger, M.J., Johnson, J.M., Rohl, C.A., van Nas, A., Mehrabian, M., Drake, T.A., Lusis, A.J., Smith, R.C., Guengerich, F.P., Strom, S.C., Schuetz, E., Rushmore, T.H., Ulrich, R.: Mapping the genetic architecture of gene expression in human liver **6**(5), 107. doi:10.1371/journal.pbio.0060107. Accessed 2015-10-19
41. Garnier, S., Truong, V., Brocheton, J., Zeller, T., Rovital, M., Wild, P.S., Ziegler, A., Munzel, T., Tiret, L., Blankenberg, S., Deloukas, P., Erdmann, J., Hengstenberg, C., Samani, N.J., Schunkert, H., Ouwehand, W.H., Goodall, A.H., Cambien, F., Trégouët, D.-A., The Cardiogenics Consortium: Genome-wide haplotype analysis of cis expression quantitative trait loci in monocytes **9**(1), 1003240. doi:10.1371/journal.pgen.1003240. Accessed 2015-10-19
42. Fairfax, B.P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., Knight, J.C.: Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression **343**(6175), 1246949. doi:10.1126/science.1246949. Accessed 2015-10-19
43. Dick, K.J., Nelson, C.P., Tsaprouni, L., Sandling, J.K., Aïssi, D., Wahl, S., Meduri, E., Morange, P.-E., Gagnon, F., Grallert, H., Waldenberger, M., Peters, A., Erdmann, J., Hengstenberg, C., Cambien, F., Goodall, A.H., Ouwehand, W.H., Schunkert, H., Thompson, J.R., Spector, T.D., Gieger, C., Trégouët, D.-A., Deloukas, P., Samani, N.J.: DNA methylation and body-mass index: a genome-wide analysis **383**(9933), 1990–1998. doi:10.1016/S0140-6736(13)62674-4. Accessed 2015-10-19
44. Johnson, A.D., Handsaker, R.E., Pulit, S.L., Nizzari, M.M., O'Donnell, C.J., Bakker, P.I.W.d.: SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap **24**(24), 2938–2939. doi:10.1093/bioinformatics/btn564. Accessed 2015-10-19
45. Westra, H.-J., Peters, M.J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M.W., Fairfax, B.P., Schramm, K., Powell, J.E., Zhernakova, A., Zhernakova, D.V., Veldink, J.H., Van den Berg, L.H., Karjalainen, J., Withoff, S., Uitterlinden, A.G., Hofman, A., Rivadeneira, F., 't Hoen, P.A.C., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A.B., Hernandez, D.G., Nalls, M.A., Homuth, G., Nauck, M., Radke, D., Völker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S.A., Enquobahrie, D.A., Lumley, T., Montgomery, G.W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R.C., Visscher, P.M., Knight, J.C., Psaty, B.M., Ripatti, S., Teumer, A., Frayling, T.M., Metspalu, A., van Meurs, J.B.J., Franke, L.: Systematic identification of trans eQTLs as putative drivers of known disease associations **45**(10), 1238–1243. doi:10.1038/ng.2756. Accessed 2015-10-19
46. Ward, L.D., Kellis, M.: HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants **40**, 930–934. doi:10.1093/nar/gkr917. Accessed 2015-10-19
47. Barenboim, M., Manke, T.: ChroMoS: an integrated web tool for SNP classification, prioritization and functional interpretation **29**(17), 2197–2198. doi:10.1093/bioinformatics/btt356. Accessed 2015-10-19
48. Boyle, A.P., Hong, E.L., Hariharan, M., Cheng, Y., Schaub, M.A., Kasowski, M., Karczewski, K.J., Park, J., Hitz, B.C., Weng, S., Cherry, J.M., Snyder, M.: Annotation of functional variation in personal genomes using RegulomeDB **22**(9), 1790–1797. doi:10.1101/gr.137323.112. Accessed 2015-10-19
49. Berditchevski, F.: Complexes of tetraspanins with integrins: more than meets the eye **114**(23), 4143–4151. Accessed 2015-10-19
50. Lau, L.-M., Wee, J.L., Wright, M.D., Moseley, G.W., Hogarth, P.M., Ashman, L.K., Jackson, D.E.: The tetraspanin superfamily member CD151 regulates outside-in integrin alpha-iib-beta3 signaling and platelet function **104**(8), 2368–2375. doi:10.1182/blood-2003-12-4430. Accessed 2015-10-19
51. Goschnick, M.W., Lau, L.-M., Wee, J.L., Liu, Y.S., Hogarth, P.M., Robb, L.M., Hickey, M.J., Wright, M.D., Jackson, D.E.: Impaired "outside-in" integrin alpha-iib-beta3 signaling and thrombus stability in TSSC6-deficient mice **108**(6), 1911–1918. doi:10.1182/blood-2006-02-004267. Accessed 2015-10-19
52. Bailey, R.L., Herbert, J.M., Khan, K., Heath, V.L., Bicknell, R., Tomlinson, M.G.: The emerging role of tetraspanin microdomains on endothelial cells **39**(6), 1667–1673. doi:10.1042/BST20110745. Accessed 2015-10-19
53. Poeter, M., Brandherm, I., Rossaint, J., Rosso, G., Shahin, V., Skryabin, B.V., Zarbock, A., Gerke, V., Rescher, U.: Annexin a8 controls leukocyte recruitment to activated endothelial cells via cell surface delivery of CD63 **5**, 3738. doi:10.1038/ncomms4738. Accessed 2015-10-19
54. Traiffort, E., O'Regan, S., Ruat, M.: The choline transporter-like family SLC44: Properties and roles in human diseases **34**(2), 646–654. doi:10.1016/j.mam.2012.10.011. Accessed 2015-10-19
55. Greinacher, A., Wesche, J., Hammer, E., Fürll, B., Völker, U., Reil, A., Bux, J.: Characterization of the human neutrophil alloantigen-3a **16**(1), 45–48. doi:10.1038/nm.2070. Accessed 2015-10-19
56. Curtis, B.R., Cox, N.J., Sullivan, M.J., Konkashbaev, A., Bowens, K., Hansen, K., Aster, R.H.: The neutrophil alloantigen HNA-3a (5b) is located on choline transporter-like protein 2 and appears to be encoded by an r>q154 amino acid substitution **115**(10), 2073–2076. doi:10.1182/blood-2009-11-248336. Accessed 2015-10-19
57. Bayat, B., Tjahjono, Y., Sydykov, A., Werth, S., Hippenstiel, S., Weissmann, N., Sachs, U.J., Santoso, S.: Anti–human neutrophil antigen-3a induced transfusion-related acute lung injury in mice by direct disturbance of lung endothelial cells **33**(11), 2538–2548. doi:10.1161/ATVBAHA.113.301206. Accessed 2015-10-19

Épidémiologie épi-génétique de biomarqueurs du risque cardiovasculaire : intérêt de l'étude de la méthylation de l'ADN à partir d'échantillons sanguins

La méthylation de l'ADN permet, via des remodelages de la chromatine et le recrutement de diverses protéines partenaires, de réguler l'expression des gènes. Des défaillances dans ces mécanismes de régulation peuvent modifier la susceptibilité individuelle face à certaines pathologies, notamment cardiovasculaires. Bien que les différents types cellulaires puissent avoir différents profils de méthylation, l'utilisation de l'ADN provenant de cellules sanguines permet de découvrir de nouveaux mécanismes physiopathologiques. Ce projet de thèse porte sur l'intérêt des analyses d'association méthylome entier comme stratégie alternative aux études d'association génome entier ("*GWAS*" en anglais) pour identifier de nouveaux déterminants moléculaires de biomarqueurs du risque cardiovasculaire. Pour cela, j'avais à ma disposition deux études épidémiologiques rassemblant 573 sujets pour lesquels les niveaux de méthylation de l'ADN issus du sang périphérique ont été mesurés par une puce à ADN de haute densité couvrant plus de 300 000 sites CpG.

Le premier travail que j'ai réalisé a consisté en une étude du méthylome sanguin pour identifier des profils de méthylation associés à l'indice de masse corporelle. Cette étude a permis d'identifier des marques de méthylation de l'ADN au sein du gène *HIF3A* dont les augmentations sont associées à une augmentation de l'indice de masse corporelle (*Lancet, 2014. 383(9933):1990-8*). Ces résultats suggèrent en outre que des perturbations de la voie métabolique du gène *HIF3A* pourraient avoir un rôle important dans la réponse biologique à l'augmentation du poids. Dans un second travail (*J Lipid Res, 2014. 55(7):1189-1191*), j'ai montré que la variabilité inter individuelle des niveaux de méthylation sanguin du gène *CPT1A* était associée à la variabilité des taux lipidiques plasmatiques. Ce travail démontre qu'il est possible de détecter à partir d'échantillons sanguins des marques de méthylation de l'ADN qui pourraient être le reflet de mécanismes épigénétiques plus spécifiques de certains types cellulaires ou de certains tissus. Le gène *CPT1A* est par exemple principalement exprimé dans le foie.

Au cours de mon travail de thèse, j'ai également étudié l'influence de la variabilité génétique sur les niveaux de méthylation de l'ADN sanguin (*Am J Hum Genet, 2015. 96(4):532-42*, *Nat Commun, 2015. 6:6326*). Cette étude a permis d'identifier près de 3 milles gènes dont les niveaux de méthylation sont associés à la présence de polymorphismes génétiques, localisés soit au sein de ces mêmes gènes (c.-à-d. effet *cis*) soit à une très grande distance (plus d'une mégabase voire sur un autre chromosome) (c.-à-d. effet *trans*). Ces résultats ouvrent de nouvelles perspectives pour mieux appréhender la régulation transcriptionnelle de diverses voies métaboliques.

Mots-Clés : Épidémiologie génétique, Méthylation de l'ADN, Analyse d'association méthylome entier (MWAS), Cardiovasculaire.

Epigenetics of cardiometabolic biomarkers through the study of DNA methylation patterns from blood samples

DNA methylation regulates gene expression by chromatin reshaping and the recruitment of various partner proteins. Dysregulation in these regulatory mechanisms can influence the individual susceptibility to some pathologies, including cardiovascular disorders. Although different cell types can have different methylation patterns, the use of DNA from blood cells has recently been proposed as an interesting tool to discover new epigenetic related pathophysiological mechanisms. This PhD project focuses on the interests of the methylome-wide association analyses as an alternative strategy to the fashion genome-wide association studies ("*GWAS*") approach to identify new molecular determinants of cardiovascular risk biomarkers. For my project, I had access to two epidemiological studies collecting together 573 subjects in which DNA methylation levels from peripheral blood cells were measured by a high density DNA microarray that covers more than 300 000 CpG sites.

The first work I conducted consisted in a study of blood methylome to identify methylation profiles associated with body mass index. This study led to the identification of DNA methylation marks at the *HIF3A* gene whose increases are associated with an increase in body mass index (*Lancet, 2014. 383(9933):1990-8*). These results suggest that a disruption of the metabolic pathway *HIF3A* gene could have an important role in the biological response to the increase of the weight. In a second work (*J Lipid Res, 2014. 55(7):1189-1191*), I showed that the inter-individual variability in *CPT1A* methylation levels in blood were associated with variability of plasma lipid levels. This work demonstrates that it is possible to detect DNA methylation marks from blood samples that could reflect epigenetic mechanisms that occur primarily in specific cells or tissues. The *CPT1A* gene is for example mainly expressed in the liver.

During my PhD, I also studied the influence of the genetic variability on the methylation levels from blood DNA (*Am J Hum Genet, 2015. 96(4):532-42*, *Nat Commun, 2015. 6:6326*). This work has identified nearly 3000 genes whose methylation levels are associated with the presence of genetic polymorphisms, located either within these same genes (ie *cis* effect) or at a very large distance (more than one megabase or to another chromosome) (ie *trans* effect). These results open new perspectives to better understand the transcriptional regulation of various metabolic pathways.

Keywords : Genetic Epidemiology, DNA methylation, Methylome-Wide Association Study (MWAS), Cardiovascular.