L'impact des variations génétiques et épigénétiques sur les réponses transcriptionnelles aux stimuli environnementaux

Lucas Husquin

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Génétique et épigénétique humaine

École doctorale nº515: Complexité du vivant

réalisée

au Laboratoire de Génétique Évolutive Humaine

sous la direction de Lluis Quintana-Murci

présentée par

Lucas HUSQUIN

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

L'impact des variations génétiques et épigénétiques sur les réponses transcriptionnelles aux stimuli environnementaux

soutenue le 26 Septembre 2019

devant le jury composé de :

M^me^	Déborah BOURCH'IS	Présidente du jury
M.	Vincent COLOT	Rapporteur
M.	Olivier DELANEAU	Rapporteur
M^me^	Raphaëlle CHAIX	Examinatrice
M.	Hugues ASCHARD	Examinateur
M.	Lluis QUINTANA-MURCI	Directeur de thèse

Remerciements

Mes premiers remerciements vont aux membres de mon jury : Déborah Bourch'is, Vincent Colot, Olivier Delaneau, Raphaëlle Chaix et Hugues Aschard. Merci à vous tous de me faire l'honneur d'évaluer cette thèse et de participer à la soutenance à venir. Une pensée particulière à Hugues qui m'a guidé pendant mes premiers pas sur les chemins de la causalité et dont les conseils ont été d'une grande aide au cours de cette thèse.

Lluis ensuite, merci de m'avoir accueilli dans ton laboratoire et de m'avoir offert l'opportunité de pouvoir conduire mes travaux de thèse dans ces conditions. Merci pour ta confiance vis à vis de toutes les analyses ayant trait à la méthylation, merci pour tes conseils et les discussions scientifiques au cours des PhD meetings hebdomadaires qui m'ont grandement aidé tout au long de ces dernières années à ne pas (trop) me disperser. Merci enfin pour les heures de travail que nous avons passées dans ton bureau à travailler ensemble sur les manuscrits de mes articles qui resteront sûrement comme les journées de travail les plus enrichissantes de cette thèse.

J'ai maintenant une pensée émue pour l'ensemble des membres du laboratoire, passés ou présents, qui ont tous contribué à l'ambiance particulière qui existe au sein de cette équipe. Dame Hélène, merci pour ton attention de tous les instants, pour m'avoir remonté le moral dans les moments durs, pour tes leçons de vie, tes conseils de lecture, et bien évidemment pour tes fameux cookies ! Maître Maximus, merci pour avoir brisé mes rêves avec autant d'abnégation, pour ton inexpugnable disponibilité et patience à la moindre de mes questions statistiques et pour ta bonne humeur inaltérable (et merci d'avance pour *The Wire*). Aux personnes avec qui j'ai partagé mon bureau, Mary et Marie en particulier, merci pour votre bonne humeur, votre méchanceté (mean office power) et d'avoir supporté mes excentricités tout ce temps ! À Martin, Jérémy, Maguelonne, Christine, Romuald, Sébas', Javier, Lara, Guillaume, Etienne, Jacob, merci pour les moments de détente, sur la terrasse, pendant les déjeuners, ou en dehors du laboratoire. Enfin, une pensée particulière pour Maud, ma mentor, qui m'a mis le pied à l'étrier quand je suis arrivé, petit stagiaire de M2 n'ayant jamais (ou presque) codé de ma vie et sans qui cette thèse aurait été bien différente.

La thèse n'étant que la conclusion d'un chemin qui a commencé bien avant, je tiens à remercier tous ceux qui ont fait un bout de route avec moi, en commençant par les amis de Grieu avec qui j'ai grandi : Charly, Lauren, Raphaël. À mon bro du collège Aksel que j'ai perdu de vue, plus pour longtemps j'espère. À tous les amis de voyage, que ce soit à Saint-Aubin, Barcelone, Budapest, Cologne, Milan, Londres ou même au Mordor : Justine et Justine, Lulu Love, Etienne, Hélène, Bianca, Alice, Victor, Clarisse et Marie. Vivement la prochaine destination ! Enfin aux copains de l'ENS, mention spéciale à Thomas, Adrien, Vincent, Victor, Solène, Martin et Florian. Un merci particulier à Florian (et Martin) pour toutes ces soirées d'aventurier ou de 7DTD qui nous ont permis de relâcher la pression. Un autre à Thomas pour l'expérience du Man, l'été dernier (c'est reparti en 2020 ?). Et

un merci encore plus particulier à Martin pour m'avoir supporté, dans tous les sens du terme, dans les bons et les mauvais moments, et pour être le meilleur coloc' qu'on puisse espérer.

J'ai gardé le meilleur pour la fin : merci à la smala, Christine, Jano, Djam', Elisa et Sarah pour avoir tant participé à mon éducation et pour ces innombrables soirées terminées à chanter autour d'un feu. Merci enfin à mes parents Isabelle et Bernard, à mon frère Sam', à ma soeur Mélo et à ma soeur d'adoption Sarah d'être la famille la plus extraordinaire et aimante qu'on puisse rêver d'avoir. Merci pour les valeurs que vous m'avez transmises et que vous continuez à me transmettre au quotidien, pour votre soutien indéfectible, et toutes vos preuves d'amour.

Table des figures

Liste des tableaux

Table des matières

Introduction

Un des enjeux majeurs de la biologie est de comprendre comment les organismes s'adaptent aux modifications de leur environnement. Y répondre permettrait d'avancer grandement dans la compréhension du monde qui nous entoure, et en particulier de mieux entrevoir comment notre espèce, en particulier, a évolué. Le développement des techniques de séquençage de l'ADN a permis d'améliorer considérablement notre compréhension du rôle des variations génétiques inter-populationnelles sur l'adaptation de l'espèce humaine. Toutefois, si ces variations génétiques permettent une adaptation à des modifications pérennes de l'environnement, elles sont moins propices à l'adaptation à des variations brusques et réversibles de notre environnement. De ce fait, il est possible de s'interroger sur l'existence d'autres mécanismes d'adaptation impliquant les marques épigénétiques, que l'on peut voir comme une seconde couche d'information qui joue le rôle d'un tampon à l'interface entre la génétique et l'environnement. En effet, on peut considérer que l'environnement exerce une influence à long terme sur la variabilité génétique, mais aussi à court terme sur la variabilité épigénétique. Malgré cela, et bien que le rôle de l'épigénétique dans la régulation de l'expression des gènes ait été largement décrit, peu d'études s'intéressent au rôle que des variations épigénétiques inter-populationnelles pourraient avoir dans notre adaptation à des modifications de notre environnement.

Au cours des 20 dernières années, il a été démontré que les pressions de sélection dues aux pathogènes sont parmi les plus fortes pressions de l'environnement auxquelles l'espèce humaine a dû faire face et que l'adaptation à de nouveaux pathogènes est responsable d'une grande part de la variabilité du génome humain. En particulier, il existe de nombreux exemples de gènes de l'immunité présentant des signatures d'adaptation locale, ce qui suggère que des variations fonctionnelles dans la séquence de ces gènes pourraient avoir donné un avantage pour la survie de certaines populations humaines. Par ailleurs, un nombre croissant d'études ont mis en lumière le rôle adaptatif des variants génétiques impliqués dans la régulation de l'expression des gènes de l'immunité. Toutefois, très peu d'études se sont intéressées au rôle de l'épigénétique dans la réponse transcriptionnelle à l'infection, ce qui permettrait pourtant de mieux comprendre les phénomènes adaptatifs de réponse de l'hôte au pathogène.

Dans ce cadre, notre hypothèse de travail initiale était que la variation épigénétique joue un rôle majeur dans la régulation de l'expression des gènes de l'immunité, et encore plus important dans le cadre de la réponse transcriptionnelle à l'activation immunitaire. Pour tester cette hypothèse, la première partie de ma thèse s'est concentrée sur l'étude des réponses transcriptionnelles de monocytes primaires à différentes bactéries et stimuli viraux, chez 200 individus d'origine européenne et africaine. Ceci m'a permis de tester si le rôle supposé de la méthylation de l'ADN dans la régulation de l'expression et de la réponse transcriptionnelle à l'infection était partagé entre populations. Dans un second temps, j'ai étudié comment l'âge et le sexe pouvaient affecter le rôle régulateur de la

méthylation dans une population génétiquement homogène. De plus, cette seconde étude m'a permis de m'intéresser à la spécificité cellulaire des interactions entre méthylation de l'ADN et expression des gènes de l'immunité.

Chapitre 1

La régulation de l'expression des gènes comme objet de la sélection naturelle

La caractérisation du vivant est une tâche ardue. Elle passe par l'attribution d'une liste plus ou moins consensuelle de propriétés à des entités physiques que l'on considère comme des organismes vivants. Parmi ces propriétés, notons la capacité des êtres vivants à protéger, exprimer et transmettre l'information contenue par les molécules d'acide désoxyribonucléiques (ADN) (Benner, 2010). Ce sont les propriétés physico-chimiques de cette molécule qui lui ont conféré son rôle central dans la genèse de l'ensemble des organismes des trois branches de l'arbre du vivant.

1.1 Description de l'organisation du génome

1.1.1 Composition de l'ADN

La molécule d'ADN est une macromolécule biologique constituée d'une répétition d'éléments de base : les nucléotides. Ces monomères, composés d'un ose, le désoxyribose, et de l'une des quatre bases azotées parmi l'adénine, la thymine, la guanine ou la cytosine, sont liés de manière covalente les uns aux autres via un phosphate. C'est cette succession de bases azotées qui détermine la séquence du brin d'ADN, et établit donc l'information génétique portée par l'ADN. Une molécule d'ADN est en général constituée de deux brins antiparallèles enroulés l'un autour de l'autre sous la forme d'une double hélice et liés l'un à l'autre par un appariement entre paires de bases azotées. Cette structure est rendue possible via l'établissement de liaisons hydrogènes, selon des règles strictes d'appariement : deux liaisons hydrogènes entre une adénine et une thymine, et trois liaisons hydrogènes entre une cytosine et une guanine (voir figure 1.1). Ces règles, qui excluent donc normalement l'appariement entre adénine et guanine, et entre cytosine et thymine, ne rendent possible l'appariement que lorsque les deux brins d'ADN sont complémentaires. Cette caractéristique permet donc une duplication de l'information génétique portée par une molécule d'ADN, et, en conséquence, la réparation d'un brin endommagé à partir de l'autre brin resté intact. De plus, la succession d'oses et de phosphates forme une structure résistante aux clivages, puisque liés de façon covalente — et donc nécessitant une quantité importante d'énergie pour rompre ces liaisons (Grandbois et al., 1999). Élucidée par Watson, Crick et Franklin en 1953 (Watson & Crick, 1953), cette structure bicaténaire hélicoïdale, plus connue sous le nom de double hélice d'ADN, confère donc une grande stabilité à l'information génétique en plaçant les bases azotées dans un environnement

stable et possédant un mécanisme de réparation en cas d'erreur. Enfin, la transmission de cette information est assurée grâce à un processus connu sous le nom de réplication de l'ADN. Ce mécanisme, au cours duquel une molécule d'ADN est dupliquée de manière presque identique — le taux d'erreur varie de 10^{-7} à 10^{-8} mutation par paire de base par réplication chez *Escherichia coli*, et pourrait même descendre jusqu'à 10^{-9} chez les eucaryotes (Schaaper, 1993 ; Loeb et al., 2003 ; McCulloch & Kunkel, 2008) — est une étape primordiale de la division cellulaire. Chaque molécule d'ADN de la cellule mère est répliquée en deux molécules d'ADN, permettant ainsi aux deux cellules filles de recevoir la même information génétique que celle portée par la cellule mère.

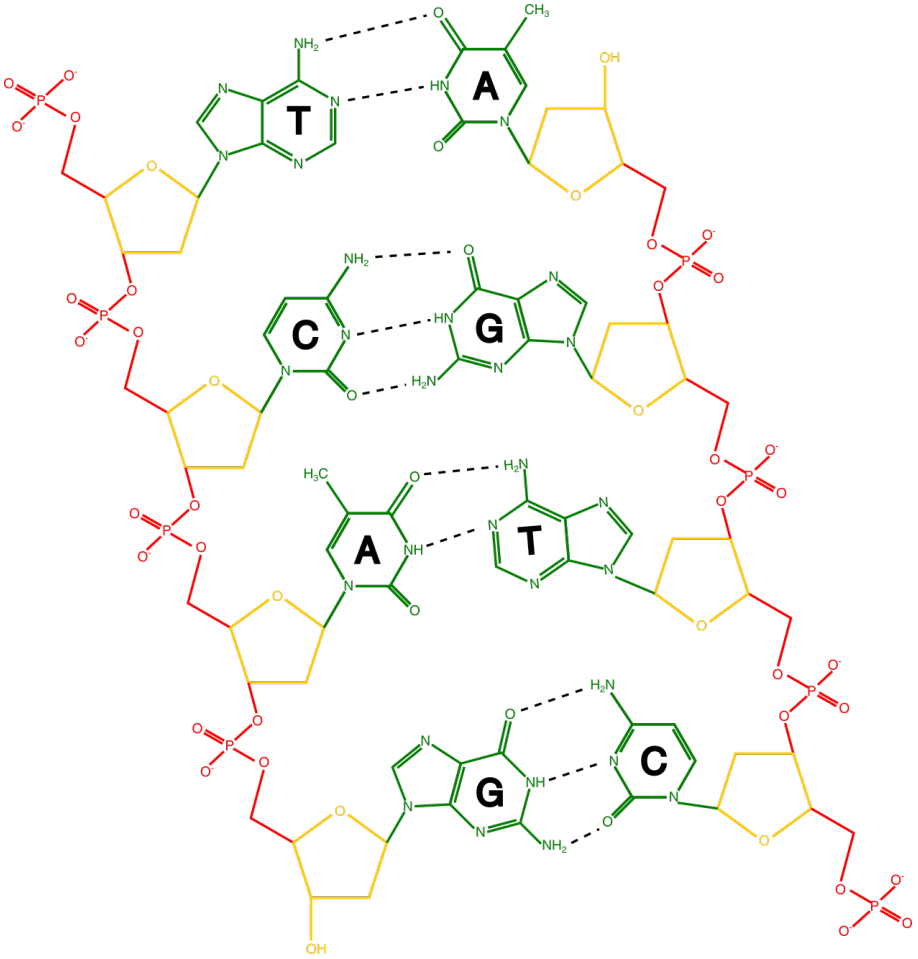

Fig. 1.1 L'appariement des bases dans l'ADN

Les ponts phosphates sont indiqués en rouge, le squelette de désoxyribose en jaune, les bases azotées en vert, et les liaisons hydrogènes permettant l'appariement en pointillés noirs.

1.1.2 Accessibilité de l'ADN

Comme nous venons de le voir, au sein d'un organisme vivant chaque cellule possède, à quelques exceptions près, une copie exacte des mêmes molécules d'ADN. Chez l'Homme, en partant d'une cellule-oeuf possédant 23 paires de chromosomes, le développement nous conduit, par divisions successives, à un corps adulte constitué de quelques milliers de milliards de cellules, chacune possédant sa propre version des 23 paires de chromosomes. Avec

un espacement estimé à 0.34 nm entre deux nucléotides successifs, il est estimé que, mises bout à bout, les paires de bases constituant le génome humain atteindraient une longueur totale de presque 2 mètres (Alberts, 2002). Pour un être humain on atteint donc une longueur totale de plusieurs dizaines de millards de kilomètres d'ADN ! Bien évidemment, l'homme moyen ne mesurant que 1.70m et la femme moyenne 1.605m (chiffres donnés à titre indicatif, tirés d'une étude de l'INSEE de 1970 en France (Valdelièvre & Charraud, 1981)), il apparaît que l'ADN ne se trouve pas dans un état libre dans les cellules. À l'échelle de la cellule, réussir à faire tenir l'ensemble de l'ADN dans un noyau d'un diamètre moyen approximant 6 µm est équivalent à essayer d'empaqueter une pelote de fil très fin long de 40 km dans une balle de tennis (Alberts, 2002). Cet exploit est rendu possible par l'action de protéines spécifiques, qui vont permettre de condenser l'ADN, dont nous discuterons maintenant quelques exemples (figure 1.2).

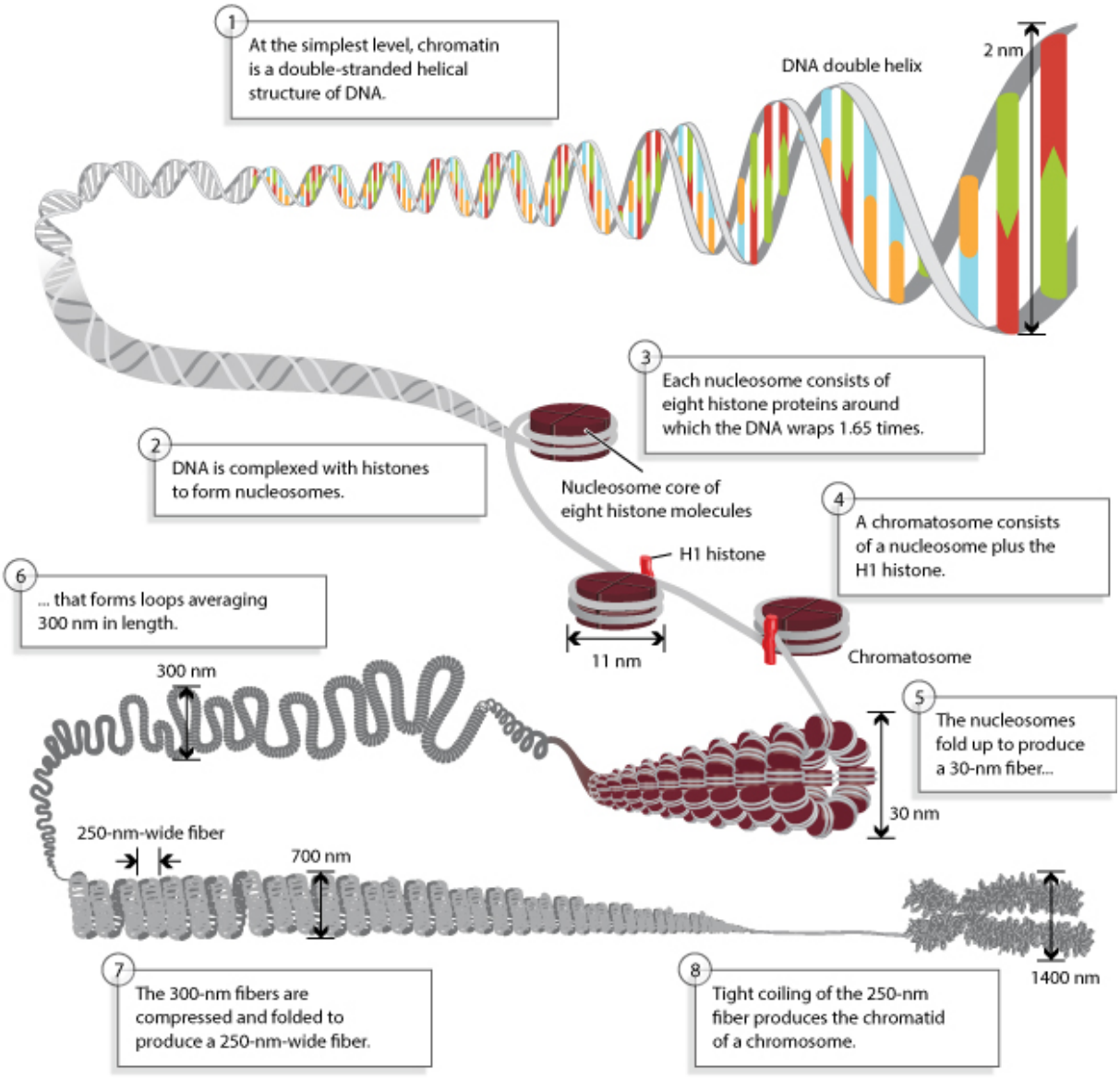

Fig. 1.2 Les différents niveaux de condensation de l'ADN
Figure tirée de (Annunziato, 2008)

En effet, plusieurs niveaux de compaction existent, le plus important et sûrement le plus connu est le chromosome mitotique, qui est la forme que prend chaque chromosome lors de la division cellulaire, avant la séparation de la cellule mère en deux cellules filles. Néanmoins, pour des soucis de clarté, intéressons-nous dans un premier temps au premier niveau de compaction de l'ADN : le nucléosome. Celui-ci implique un noyau protéique, sous la forme d'un octamère d'histones composé de deux exemplaires de chacune des histones H2A, H2B, H3 et H4, autour duquel environ 146 paires de bases de l'ADN s'enroulent sur 1.65 tours, pour une longueur totale de 11 nm (Luger et al., 1997 ; Wolffe, 1998). Cette structure nucléoprotéique, qui se répète toutes les 200 $\pm$40 pb tout au long du génome des eucaryotes, permet d'atteindre un premier niveau de compaction de l'ADN qui est sous cette forme six fois plus compact que la même longueur d'ADN nu. Comme nous venons de le mentionner le nucléosome est une structure répétée le long du génome, cette répétition est entrecoupée de segments d'ADN dits de "liaison", d'une longueur qui varie entre 25 et 70 paires de bases selon l'espèce considérée, et stabilisée par l'histone H1 (Bednar et al., 1998 ; Woodcock, 2006). Ceci conduit à l'établissement d'une structure d'un ordre supérieur, la chaîne de 30nm, dont l'architecture et les modalités de compaction sont encore incertaines, mais qui permet d'atteindre un facteur de compaction estimé entre 30 et 40 (Robinson et al., 2006 ; Grigoryev & Woodcock, 2012). Enfin, le dernier niveau de condensation de l'ADN est la conséquence du surenroulement de la chaîne de 30 nm que nous venons de décrire, et de l'action des protéines SMC (pour *Structural Maintenance of Chromosome*) (Vologodskii & Cozzarelli, 1994 ; Gassmann et al., 2004). Ce surenroulement qui peut être positif (dans le sens de l'enroulement de la double hélice) ou négatif (dans le cas contraire) permet, dans sa forme la plus extrême, l'établissement des chromosomes mitotiques précédemment évoqués. Pour atteindre ce niveau de compaction (1 :10,000), des protéines SMC telles que les cohésines et condensines, sont requises (Gassmann et al., 2004 ; M. Sun et al., 2011).

L'ensemble des niveaux de compaction que nous venons de décrire nécessite une interaction entre ADN et protéines, la structure résultant de cette interaction est appelée chromatine et correspond à l'état sous lequel l'ADN est trouvé dans le noyau des cellules eucaryotes. On distingue deux formes de chromatine, l'euchromatine d'une part, formée par le simple enroulement de l'ADN autour des coeurs d'histones et l'hétérochromatine de l'autre, plus généralement formée par les fibres de 30 nm et donc dans un état plus condensé que l'euchromatine. Nous venons ici d'établir comment les molécules d'ADN portent une information, la protègent et la transmettent d'une génération à l'autre de manière pérenne. Afin de compléter ce qui fait de l'ADN la molécule essentielle au vivant, intéressons-nous maintenant à l'expression de l'information génétique.

1.2 Régulation de l'expression de l'ADN

Comme nous venons de le mentionner, les molécules d'ADN sont essentiellement une succession de nucléotides formant un code et portant donc une information. En particulier certaines parties de l'ADN, plus communément appelées gènes, portent une succession de nucléotides, qui lus trois par trois par les organismes vivants peuvent être interprétés en acides aminés, les briques unitaires des protéines. Cette correspondance entre trios de nucléotides (plus connus sous le nom de "codons") et acides aminés est appelée le code génétique et est à la base de la capacité qu'ont les organismes vivants à interpréter l'information contenue par les séquences d'ADN (Crick et al., 1961). Ce processus de lecture de

l'information génétique est appelé expression de l'ADN, ou expression génique, et repose sur deux étapes fondamentales : la transcription de l'ADN en ARN messager (ARNm), puis la traduction du-dit ARNm en protéine (nous nous restreindrons ici au cas des gènes codant des protéines, bien qu'il existe aussi de nombreux exemples de gènes codant des petits ARN circulants, longs ARN non-codants ou autres (Esteller, 2011 ; Uszczynska-Ratajczak et al., 2018)). La première de ces étapes est intimement liée à l'accessibilité de l'ADN puisque nécessitant un accès direct à la séquence de l'ADN par toute une machinerie de protéines requise pour la transcription en ARNm. De ce fait, on comprend facilement le rapport important qui existe entre la transcription et le taux de compaction de l'ADN : si l'euchromatine concerne en général les gènes activement transcrits, l'hétérochromatine regroupe les zones non-actives de l'ADN. Notons qu'on distingue l'hétérochromatine constitutive, commune à l'ensemble des cellules, de l'hétérochromatine facultative quand spécifique au type cellulaire et/ou à l'état de différenciation. Dans cette partie nous allons nous intéresser à l'ensemble des mécanismes qui permettent la régulation de l'expression de l'ADN, avec dans un premier temps une description des éléments régulateurs du génome. Pour des soucis de clarté, nous nous concentrerons dorénavant sur le cas de l'être humain.

1.2.1 Les différents éléments régulateurs

Le séquençage du génome humain, complété depuis bientôt 20 ans, a grandement contribué à l'identification de l'ensemble des régions fonctionnelles de l'ADN (Lander et al., 2001 ; Venter et al., 2001). Ces régions, généralement non transcrites, jouent un rôle primordial dans l'établissement des profils spatiaux et temporels d'expression, que ce soit au cours du développement ou de la vie d'un individu. Bien que l'expression des gènes puisse être régulée à différents niveaux, pendant ou après la transcription des ARNm, il est considéré que la majorité des processus de régulation se produisent au cours de l'initiation de la transcription. Chez l'Homme, c'est l'ARN polymérase II, couplée à différents facteurs de transcription généraux (GTF), qui est responsable de la transcription des gènes codant des protéines, cas sur lequel nous nous concentrerons ici (Orphanides et al., 1996). Pour ces gènes, il existe deux grandes familles de régions régulatrices en *cis* : (i) les promoteurs, parmi lesquels on distinguera les promoteurs principaux et les éléments régulateurs proximaux et (ii) les éléments de régulation distale, qui comprennent les *enhancers*, *insulators*, *silencers* ou encore les régions de contrôle de locus (LCR). Les promoteurs principaux correspondent à la région d'ADN au début du gène où le complexe protéique formé par l'ARN polymérase II et ses co-facteurs va s'arrimer. Ils définissent la position du TSS (pour *Transcription Start Site*, que l'on peut traduire en "site de démarrage de la transcription"), ainsi que le sens de la transcription (Smale & Kadonaga, 2003). Cet arrimage de la machinerie de transcription est rendu possible par la reconnaissance de certaines séquences spécifiques d'ADN, telles que la TATA box qui fût la première découverte dans les années 1980, ou l'élément motif 10, identifié plus récemment (Motif Ten Element, MTE) (C. Y. Lim et al., 2004). Ces promoteurs principaux sont le plus généralement associés à des éléments régulateurs proximaux, qui sont des régions situées au maximum à quelques centaines de bases en amont du promoteur et permettent la fixation de nombreux activateurs requis pour la transcription. Dans ces quelques centaines de bases en amont du TSS d'un gène donné on trouve en général plusieurs régions régulatrices de ce type, régulant de manière synergique l'expression du gène en question (Lonard

& O'Malley, 2005).

Les *enhancers* sont des régions ressemblant très fortement aux éléments régulateurs proximaux que nous venons de décrire, à tel point que ces derniers pourraient être considérés comme des *enhancers* proximaux, dans le sens où la différence majeure réside dans le fait que les *enhancers* peuvent être situés à une distance bien plus importante du promoteur principal. Ainsi, certains *enhancers* peuvent être situés à plusieurs centaines de milliers de bases en amont du TSS, voire en aval du promoteur au sein d'un intron du gène ou même en aval du gène lui-même (Blackwood & Kadonaga, 1998). Constitués d'un groupe de sites de fixation pour facteurs de transcription, généralement proches les uns des autres, les *enhancers* permettent de contrôler la transcription de manière spécifique en fonction du tissu, ou de la période de la vie, du développement ou même du jour (M. Levine et al., 2014 ; Zabidi et al., 2015). Le mode d'action proposé pour faire rentrer en contact les facteurs se fixant sur les *enhancers* et la machinerie de transcription fixée au promoteur principal est en général un repliement de l'ADN pour former des boucles rapprochant ainsi spatialement des régions initialement éloignées les unes des autres, illustrant une nouvelle fois le lien fort existant entre la structure tridimensionnelle de l'ADN et activité transcriptionelle (Bulger & Groudine, 2011). Si les *enhancers* sont en général associés à l'activation de la transcription, d'autres types d'éléments régulateurs ont un rôle répressif, ou de contrôle de l'expression, c'est le cas des *silencers* et des *insulators*. Les *silencers* sont des régions ressemblant fortement aux *enhancers* dans leur architecture et leur position, mais attirent des répresseurs de la transcription plutôt que des activateurs. Ces répresseurs vont ensuite agir en bloquant la fixation d'activateurs ou autres facteurs de transcription, empêchant ainsi le bon déroulement du processus de transcription (L. Li et al., 2004 ; M. B. Harris et al., 2005). Les *insulators* (ou isolateurs en français), comme leur nom l'indique permettent de bloquer l'effet d'une région régulatrice sur un gène lorsque intercalés entre les deux, en particulier en bloquant la communication entre *enhancer* et promoteur. Une autre propriété importante des isolateurs est de stopper la propagation de l'hétérochromatine, en formant une barrière à celle-ci. Longues de 500 à 3,000 bases, ces régions sont intimement liées à l'activité du facteur de transcription CTCF, qui pourrait être responsable du bloquage de la communication entre *enhancer* et promoteur via la structure tridimensionnelle prise par l'ADN quand CTCF s'y fixe (Splinter et al., 2006 ; Herold et al., 2012 ; Kim et al., 2015). Finalement, les LCR sont des groupes des différents éléments régulateurs que nous venons de décrire et dont le rôle est la régulation de la transcription d'un locus entier, ou d'un cluster de gènes, de manière spécifique dans un tissu ou à un moment précis, ici aussi en favorisant le rapprochement de régions d'ADN au sein d'un même chromosome, mais aussi provenant de deux chromosomes différents (Dean, 2006).

1.2.2 Facteurs environnementaux et intrinsèques de la variabilité de l'expression génique

Bien qu'il existe des différences dans les profils d'expression entre individus, différences qui se renforcent plus les individus comparés appartiennent à des populations d'ascendance lointaine (et donc présentant de fortes divergences génétiques, ce que nous étudierons plus en avant dans la partie suivante) et/ou vivant dans des environnements fortement différenciés, l'étude des facteurs influençant les profils d'expression représente de nombreux défis. Entre un individu venant d'une population A et vivant dans un environnement A', et

un individu d'une population B et dans un environnement B', l'attribution des différences d'expression à la génétique ou à l'environnement est une tâche qui est rendue ardue par l'impossibilité (ou la très grande difficulté) de quantifier une variable aussi multi-factorielle que l'environnement. En effet, le tabagisme, l'alimentation, l'exercice, la température, les facteurs sociaux-économiques ou l'exposition à des polluants entre autres sont autant de facteurs qui vont définir l'environnement global. Ceci est d'autant plus vrai que bien souvent ces variables sont aussi inter-connectées, et que certaines pratiques culturelles (et donc liées à l'appartenance à une population ou l'autre) auront une influence sur certaines des variables décrites (telles que l'alimentation par exemple).

Malgré ces difficultés, il existe de nombreuses méthodes visant à homogénéiser les populations d'étude pour estimer l'effet de certaines variables en contrôlant au maximum l'effet d'autres variables, en faisant ressembler l'environnement A' à B' par exemple. Ainsi, il a été montré que de très nombreux facteurs environnementaux ont un effet sur l'expression, tels que le tabagisme (Charlesworth et al., 2010 ; Paul & Amundson, 2014), l'alimentation (Cameron-Smith et al., 2003 ; Bouchard-Mercier et al., 2013), la pollution de l'air (Chu et al., 2016 ; Fave et al., 2018) ou encore en particulier l'exposition à des pathogènes (Bootsma et al., 2007 ; Smale, 2012 ; C. Yang et al., 2016 ; Quach et al., 2016). Ces quelques exemples ne représentent que la pointe de l'iceberg vis à vis des différents facteurs environnementaux pour lesquels un impact sur l'expression a été identifié, mais donnent une bonne idée de l'importance globale de l'environnement sur l'activité transcriptionnelle. Néanmoins, ces résultats ayant été obtenus dans de nombreux cas dans des tissus hétérogènes, il est nécessaire de rester mesuré dans l'interprétation faite de telles associations. En effet, il est bien connu que ces différents facteurs environnementaux peuvent avoir un impact important sur l'hétérogénéité cellulaire des tissus, et du sang en particulier - qui reste le tissu d'étude privilégié de ce genre d'études de par son accessibilité (Huang, 2009 ; Boutens & Stienstra, 2016). De ce fait, les associations observées entre variables environnementales et activité transcriptionnelle ne pourrait être que le reflet des variations des proportions des différents types cellulaires du tissu d'étude. Enfin, il a aussi été montré que certaines variables intrinsèques telles que l'âge, ou le sexe ont un fort impact sur les profils d'expression (Grath & Parsch, 2016 ; Piasecka et al., 2018 ; S. E. Harris et al., 2017).

1.2.3 Facteurs génétiques de la variabilité de l'expression génique

Une grande partie des éléments régulateurs du génome que nous avons décrits jusqu'ici ont un mode d'action qui repose sur la reconnaissance de séquences spécifiques de l'ADN par des facteurs de transcription, que ceux-ci soient des activateurs ou des répresseurs. Or, bien que généralement conservée au sein d'une même espèce la séquence d'ADN peut subir certaines modifications, en particulier on distinguera les substitutions d'un acide nucléique par un autre nommées SNP (pour *Single Nucleotide Polymorphism*). Ce type de mutation est le plus fréquent dans le génome humain, on estime qu'entre deux individus pris au hasard on observera une moyenne d'un SNP par 1,000 paires de bases (Lander et al., 2001 ; Venter et al., 2001 ; Genomes Project et al., 2010). De plus, on évalue à 10,000 le nombre de SNP dans le génome d'un individu entre ses différentes cellules (Genomes Project et al., 2010). On comprend donc aisément que cette forte variabilité génétique, individuelle ou inter-individuelle, si elle touche les différents éléments régulateurs puisse avoir un impact

sur la fixation des facteurs de transcription, et en conséquence influer sur l'expression des gènes sous contrôle de la région régulatrice mutée.

Nécessitant plusieurs centaines d'individus, l'identification de ces variants génétiques, plus connus sous le terme d'eQTL (*expression Quantative Trait Loci*, en anglais), est l'objet d'un nombre croissant d'études depuis l'essor des méthodes de séquençage à haut débit dans la fin des années 1990 (Yan et al., 2002 ; Cheung et al., 2003 ; Monks et al., 2004 ; Spielman et al., 2007 ; C. Lee, 2018). Les régions régulatrices ayant pour certaines un effet spécifique au tissu, ces variants génétiques peuvent avoir un effet spécifique au tissu aussi, dans le cadre d'une mutation d'un *enhancer* par exemple ou un effet global, si la mutation touche par exemple le promoteur principal. Pour déterminer cette spécificité des eQTL, des consortiums, tel que GTEx (pour *Genotype-Tissue Expression*) en particulier, ont établi les profils d'expression d'un grand nombre de tissus chez un grand nombre d'individus, mettant à disposition de la communauté scientifique une base de données des variants génétiques impliqués dans la régulation de l'expression génique (Consortium, 2015 ; Võsa et al., 2018). Ces formidables outils, accessibles par tous, contribuent à une meilleure compréhension globale des processus de régulation de l'activité transcriptionnelle au sein des différentes cellules humaines.

1.3 La sélection naturelle

Les molécules d'ADN, comme nous l'avons vu, portent l'information nécessaire au développement des organismes qui joueront le rôle de "véhicules" pour les "réplicateurs" qu'elles sont. Cette théorie, dite du "gène égoïste", développée en 1976 par Richard Dawkins dans son livre *The Selfish Gene* place le gène comme élément central de la théorie de l'évolution de Charles Darwin (Dawkins, 1976). Il existe de nombreuses autres tentatives d'interprétation et d'adaptation aux connaissances actuelles de la théorie de Darwin, et toutes découlent d'une même volonté commune : comprendre comment, à partir d'un ancêtre commun unique s'est développée l'immense biodiversité observable sur notre planète aujourd'hui. En effet, deux millions d'espèces de plantes, animaux et microbes ont été identifiées à ce jour par la communauté scientifique (B Larsen et al., 2017 ; Mora et al., 2011), et il est estimé qu'il reste encore des millions, voire des milliards, d'espèces à découvrir et identifier, des centaines d'espèces nouvelles étant découvertes chaque jour (B Larsen et al., 2017 ; Scotland et al., s. d.). C'est en partie pour répondre à cette question, et trouver une alternative au fixisme - l'idée généralement acceptée jusqu'il y a quelques centaines d'années que l'ensemble des espèces vivantes existent telles qu'elles ont été créées par Dieu - que des scientifiques tels que Jean-Baptiste de Lamarck et Charles Darwin ont développé des théories que nous allons détailler dans un premier temps, avant de préciser les mécanismes généralement acceptés de nos jours par lesquels les organismes vivants évoluent.

1.3.1 Lamarckisme

C'est au XVIII^{e} siècle que des observations sur les similitudes existant entre l'anatomie de divers animaux ont donné lieu aux prémices de la science de l'évolution, en introduisant l'idée d'une filiation entre les espèces. La généralisation de cette idée en une théorie expliquant la diversité du monde vivant est faite par Lamarck au début du XIX^{e} siècle ;

ainsi naît le transformisme.

Pour Lamarck c'est la reprise de la collection d'invertébrés du Museum de Paris et l'étude de fossiles de mollusques qui le convainc que les phénomènes de divergence des espèces est le résultat de transformations successives permettant aux organismes vivants de s'adapter à leur environnement (Lamarck, 1809 ; Mayr, 1982). Basée sur l'idée de l'hérédité des caractères acquis, Lamarck développe une théorie pour expliquer le processus de l'évolution qui implique l'adaptation des organismes vivants à leur environnement immédiat. Cette adaptation induit l'aquisition de nouvelles caractéristiques (Première Loi) qui seront transmises à la descendance (Deuxième Loi) (Lamarck, 1809). Ainsi, de génération en génération les organismes vivants se transforment, et à terme donnent naissance à de nouvelles espèces qui auront évolué en s'adaptant au mieux à leur milieu. Bien que Lamarck ne fut pas le premier à concevoir l'hérédité des caractères acquis, l'origine de cette idée pouvant être tracée aussi loin qu'Aristote, il fut le premier à percevoir l'impact de ce concept sur l'évolution des espèces. En particulier, même si pour beaucoup d'autres scientifiques il était concevable que les individus puissent hériter des caractéristiques acquises par leurs parents, il était en revanche inimaginable que cela puisse conduire à la création de nouvelles espèces (Leroy, 1802). Lamarck, au contraire, voyait en l'hérédité des caractères acquis "un agent de changements illimités" et tira tout au long de sa vie beaucoup de satisfaction à avoir été le premier à comprendre “l'importance de cette loi et la lumière qu'elle apporte aux causes qui ont conduit à cette stupéfiante diversité du vivant" (Burkhardt, 2013). Toutefois, Lamarck n'a jamais cherché à identifier les mécanismes de transmission d'une génération à l'autre des caractéristiques, et n'a finalement que repris une idée déjà généralement admise.

1.3.2 Darwinisme et Néo-Darwinisme

Ce point, déjà problématique chez Lamarck, n'a pas été résolu par l'apport de Darwin lors de la publication de *l'Origine des espèces*, qui part du même présupposé d'une hérédité des caractères. Toutefois, il propose une explication tout à fait différente à l'évolution des espèces basée sur un mécanisme original : *la sélection naturelle* (Darwin, 1859).

Darwin a été inspiré par les travaux de Thomas Malthus qui décrivait comment dans nombre d'espèces, les naissances étaient trop nombreuses par rapport à la capacité d'accueil du milieu. Ce constat induit que seule une partie des individus mis au monde arriveront à survivre et à atteindre à leur tour l'âge de procréer et donner vie à la génération suivante. Il y a donc une *lutte pour la vie* entre individus d'une même espèce, mais aussi, comme Darwin le comprit, une lutte entre espèces dans le sens où des individus d'espèces différentes peuvent se retrouver en compétition si dépendantes d'une ressource commune, disponible en quantité limitée (J. Bowler, 1976). Par ailleurs, l'expérience de son voyage sur le *Beagle* fût essentielle dans l'élaboration de sa théorie de l'évolution. En particulier ses observations des pinsons des Galápagos, qu'il pensait être des variations d'une même espèce mais furent identifiées comme des espèces différentes par John Gould, lui donna l'idée que des variations au sein d'une même espèce puissent à terme donner naissance à des espèces différentes (Steinheimer, 2004). Ces deux idées, d'une variabilité innée associée à une lutte pour la vie, imposée par le milieu, sont les deux piliers de *la sélection naturelle*. La grande différence distinguant la vision de Darwin de celle de Lamarck (illustrée dans la figure 1.3) est que Darwin perçoit que les caractéristiques transmises d'une génération à l'autre sont *innées*, alors que selon Lamarck elles sont *acquises*.

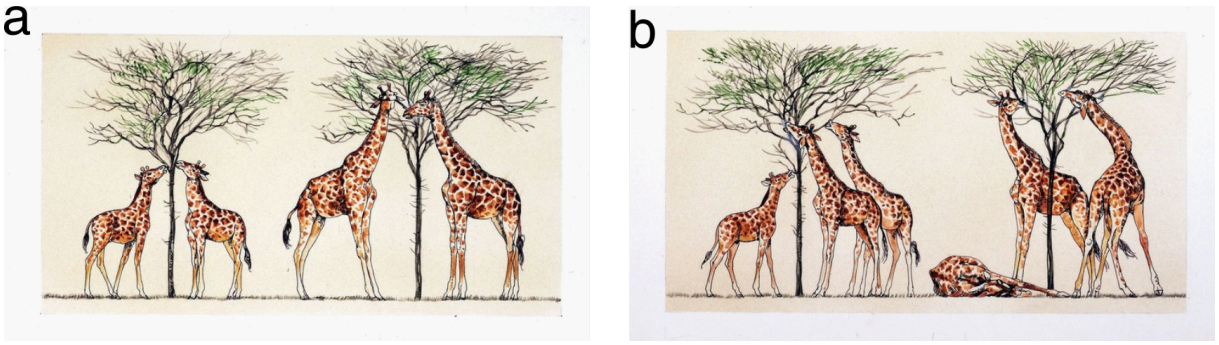

Fig. 1.3 L'évolution de la taille du cou des girafes
a) Selon Lamarck, le cou des girafes s'accroît au cours de leur vie, en réponse aux pressions de leur environnement. b) Selon Darwin, les girafes ont de manière innée un cou de longueur variable, ce qui conduit à une meilleure survie des individus ayant un cou plus long que la moyenne. ©DEAGOSTINI/LEEMAGE.

Toutefois, ni l'un ni l'autre n'ont pu proposer à l'époque une explication sur les mécanismes de la transmission de ces caractères. Darwin suggéra, faute de mieux, que chaque trait hérité est la moyenne des traits des deux parents et c'est sur ce point que la théorie de Darwin fut le plus vivement critiquée, en particulier par Fleeming Jenkin qui constata qu'un tel mécanisme de transmission conduirait inéluctablement à une homogénéisation de la population. Selon ce dernier, toute caractéristique déviant trop fortement de la norme au sein d'une population verrait son effet être absorbé à cause de ce processus d'homogénéisation propre au mode de transmission envisagé par Darwin (Jenkin, 1867). La solution à cet écueil majeur était pourtant disponible du vivant de Darwin : dès 1865, Gregor Mendel, un moine vivant à Brno, avait proposé que dans les espèces sexuées chaque individu était contrôlé par deux versions des mêmes facteurs hérités pour moitié de la mère et pour moitié du père (Mendel, 1866). Cette idée, trop révolutionnaire pour l'époque, ne fût acceptée qu'au début du XX^{e} siècle lorsque les progrès techniques rendirent possible l'observation de ce qui fut plus tard identifié comme les chromosomes. Cette redécouverte des lois de Mendel permit l'élaboration d'une nouvelle théorie combinant la sélection naturelle avec la transmission génétique : Le néo-darwinisme ou "théorie synthétique de l'évolution".

1.3.3 La diversité phénotypique et génétique comme cible et acteur de la sélection

Cette reformulation de la théorie de Darwin implique que ce ne sont donc pas les caractères observés par Darwin et ses confrères qui évoluent, mais bien le patrimoine génétique d'une population. Il est donc important de faire la distinction entre la diversité phénotypique et la diversité génétique et leurs rôles respectifs dans le cadre de la sélection naturelle. Au sein d'une population donnée, la diversité phénotypique représente la variation des caractères observables et/ou quantifiables alors que la diversité génétique correspond aux variations de la séquence d'ADN entre individus. Que ce soit dans une vision Lamarckiste, Darwiniste ou néo-Darwiniste, la théorie de l'évolution repose sur la survie des individus les plus adaptés à leur environnement. C'est donc sur la diversité phénotypique, qui fournit une variabilité plus ou moins importante d'une multitude de

caractères différents, que la sélection naturelle va avoir prise (Karlen & Krubitzer, 2006). En effet c'est l'ensemble des caractères d'un individu, autrement dit son phénotype, qui sera primordial pour déterminer sa survie et *in fine* la transmission de ses caractères. Toutefois, si l'on peut convenir que la diversité phénotypique est la cible de la sélection, il faut admettre qu'une telle diversité n'est que le résultat de l'héritage des caractères des parents via la transmission de la moitié du patrimoine génétique du père et de la mère. Ce constat amène donc à considérer les variations génétiques résultant du brassage inhérent à une telle transmission comme les acteurs principaux de la sélection naturelle. En effet, si ce sont les phénotypes qui sont sélectionnés par leur capacité à s'adapter et à survivre à un environnement donné, ce sont bien les séquences d'ADN sous-jacentes qui seront transmises (ou tout du moins des copies fidèles de ces séquences) aux générations suivantes. Nous revenons ici à l'idée d'un "réplicateur" qui serait l'unité de base de la sélection naturelle et que Dawkins a défini comme étant le gène (Dawkins, 1976 ; Lloyd, 2001). La variation en fréquence des différents gènes dans le patrimoine génétique d'une population permettrait ainsi de modifier sa diversité phénotypique, qui, ciblée par la sélection naturelle évoluera au cours des générations pour favoriser l'émergence de phénotypes mieux adaptés à l'environnement.

Dans la prochaine partie, nous nous intéresserons aux liens qui existent entre les gènes et les phénotypes et que nous venons rapidement d'aborder. Nous verrons comment la variabilité des profils d'expression au cours du temps, au sein d'un individu ou entre individus d'une même espèce ou non peut être reliée aux processus de sélection naturelle que nous venons de décrire. Ceci nous permettra de nous éloigner de la vision centrée sur le gène que nous venons d'évoquer et de considérer l'importance des mécanismes de régulation de l'expression génique dans le cadre de l'évolution.

1.4 Les phénotypes, conséquences directes de l'expression de l'ADN

L'ensemble des régions régulatrices que nous avons précédemment décrites est essentiel à l'établissement des différents patrons d'expression (spatiaux et temporels) nécessaires au développement, mais aussi à l'adaptation constante à un environnement fluctuant auquel nos cellules sont soumises (Kornberg & Tabata, 1993 ; Tung & Gilad, 2013 ; LaFlamme, 2014 ; Fave et al., 2018). Ceci provoque une très forte variabilité de l'expression génique, et ce, entre individus mais aussi au sein même d'un individu entre les cellules composant les différents tissus de son organisme (Raj & van Oudenaarden, 2008 ; Chalancon et al., 2012 ; Consortium, 2015 ; Mele et al., 2015 ; Quach et al., 2016 ; Nedelec et al., 2016 ; Breschi et al., 2016 ; Carcamo-Orive et al., 2017). Ces différences inter-cellulaires sont majoritairement le fruit de la mise en place de différents patrons d'expression au cours du développement, ce à quoi nous allons nous intéresser maintenant.

1.4.1 Établissement des profils d'expression au cours du développement

Comment les êtres vivants pluri-cellulaires tels que l'Homme mettent-ils en place les différents types cellulaires constituant leur organisme à partir d'une cellule oeuf unique ? Comment ces différents types cellulaires peuvent-ils être si différents dans leur structure et

leurs fonctions, alors même qu'ils partagent tous la même information génétique, issue de la fusion de deux gamètes lors de la fécondation (Lappin et al., 2006) ? Ces questions ont guidé les travaux de recherche de nombre de biologistes du développement au cours du XX^{e} siècle et ont mis en lumière l'importance des mécanismes de régulation de l'expression, en particulier via l'identification du rôle des gènes homéotiques (Lewis, 1978). Ces gènes, à qui l'on attribue la fonction d'architecte du développement, sont des gènes codant des facteurs de transcription largement conservés au cours de l'évolution. En effet on retrouve des homologues de ces gènes dans des organismes aussi variés que les échinodermes, les insectes, les mammifères ou les plantes (Popodi et al., 1996 ; Hirth et al., 1998 ; Theissen, 2001 ; Young et al., 2009). La famille la plus connue de gènes homéotiques est certainement la famille des gènes *Hox*, présents dans l'espèce humaine au nombre de 52 répartis dans 14 familles de gènes (Holland et al., 2007). Répartis en cluster de gènes le long du génome, ces gènes vont être exprimés dans un ordre précis (du gène en 3' du cluster aux gènes situés en 5'), permettant l'établissement de groupes de cellules destinés à former des organes ou membres spécifiques (figure 1.4). Ces domaines fonctionnels sont déterminés par l'expression de combinaisons de gènes *Hox* le long de l'axe antéro-postérieur, dont l'activité va permettre l'activation de patrons d'expression déterminants dans la réalisation du phénotype spécifique au type cellulaire concerné (Kmita & Duboule, 2003 ; Mallo & Alonso, 2013). Il existe de nombreux exemples de maladies développementales causées par des mutations touchant cette famille de gènes, telles que le développement de la synpolydactylie (Quinonez & Innis, 2014). Cette malformation des mains conduit à la fusion de certains doigts ainsi qu'au développement de doigts supplémentaires et est causée par une mutation du gène *HOXD13* (Lappin et al., 2006). Plus récemment, un certain nombre de mutations dans ces gènes ont été liées avec une prédisposition accrue à des cancers de différents type tels que des cancers du colon, des poumons ou de la prostate (Bhatlekar et al., 2014).

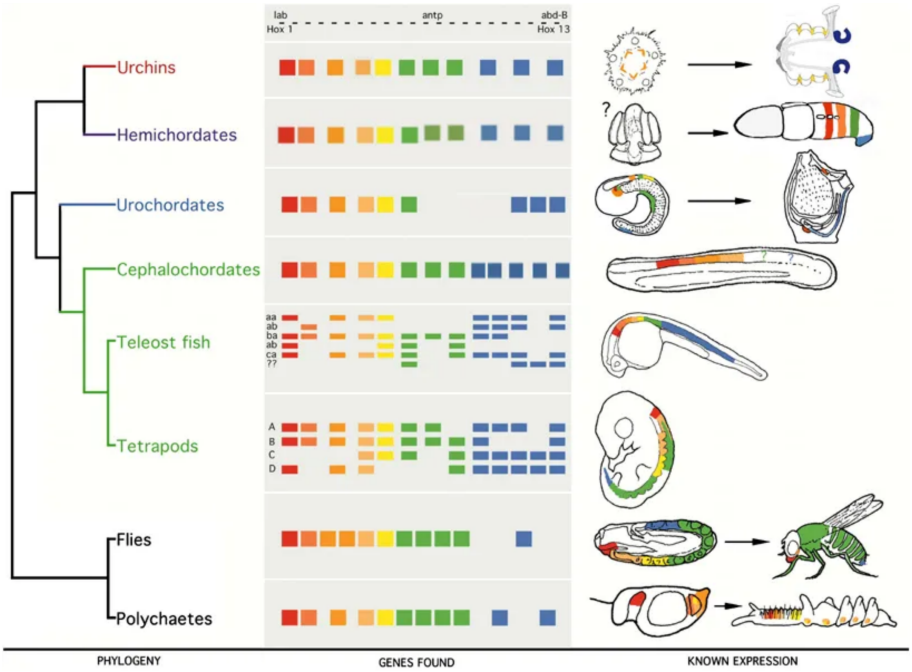

Fig. 1.4 Patrons d'expression des gènes *Hox* dans les embryons des Deutérostomiens

Figure tirée de (Swalla, 2006)

Nous venons de voir ici comment un petit nombre de gènes permet, via leur rôle régulateur de l'expression de réseaux de gènes sous-jacents, l'établissement du plan d'organisation d'organismes extrêmement complexes. Cette action coordonnée de quelques facteurs de transcription permet l'établissement de phénotypes sains, ou malades dans le cas de mutations affectant un ou plusieurs de ces gènes homéotiques. Nous avons ici un premier exemple de l'importance de la régulation de l'expression des gènes dans l'établissement du phénotype d'un individu. Cette régulation spatio-temporelle permet d'établir des patrons d'expression spécifiques aux différents types cellulaires d'un organisme, nous allons maintenant voir qu'il existe aussi une variabilité interindividuelle et inter-espèce des profils d'expression.

1.4.2 Variabilité inter-espèce et interindividuelle des profils d'expression

Comme nous l'avons vu précédemment, de nombreux facteurs environnementaux, intrinsèques et génétiques ont été identifiés pour leur impact sur l'expression de certains gènes. Il doit ainsi exister une forte variabilité des profils d'expression d'un individu à l'autre, conséquence de la variabilité génétique et/ou des différences d'environnement entre individus. En effet, de nombreuses études consacrées à l'analyse de la variabilité des profils d'expression ont été menées depuis l'avènement de techniques telles que le séquençage d'ARN qui a permis de quantifier les profils d'expression pour un grand nombre d'individus (Conesa et al., 2016).

Il est important ici de distinguer la variabilité de l'expression, entre individus au sein d'une même population, de l'expression différentielle qui implique une différence d'expression moyenne entre individus appartenant à deux populations distinctes (J. Li et al., 2010). En particulier chez l'espèce humaine, de nombreux chercheurs se sont intéressés à identifier les différences d'expression existant entre populations, tandis qu'un intérêt bien moindre était porté à l'étude de la variabilité de l'expression au sein d'une population homogène (Storey et al., 2007 ; J. Li et al., 2010). Bien que jusqu'à 83% des gènes étudiés aient été identifiés comme variables entre individus (Storey et al., 2007), une autre étude sur un nombre considérablement plus élevé d'individus européens et africains n'a pas trouvé de différences massives dans le niveau de variabilité entre individus d'une population à l'autre (J. Li et al., 2010). Ceci indique soit que l'expression de la plupart des gènes testés est sujette aux mêmes niveaux de contrainte dans les deux populations, soit que les déterminants génétiques de l'expression de ces gènes n'ont pas divergé de manière significative. L'analyse des différences d'expression entre populations humaines a été l'objet de nombreuses études, dans le but en particulier d'enrichir notre compréhension des processus d'adaptation à des environnements variés. Si les profils d'expression ne permettent pas d'isoler les différentes populations étudiées aussi bien que ne le ferait l'étude de la variabilité génétique entre populations, il apparaît néanmoins qu'un certain nombre de gènes montrent de très fortes différences d'expression entre paires de populations (Spielman et al., 2007 ; W. Zhang et al., 2008 ; Stranger et al., 2012 ; Quach et al., 2016). De plus, l'identification des variants génétiques responsables des différences d'expression entre populations, conduite dans la plupart des exemples cités, permet d'identifier des régions du génome ou des fonctions biologiques potentiellement sélectionnées au cours de l'histoire évolutive de ces populations, telles que le système immunitaire (Quach et al., 2016 ; Nedelec et al., 2016 ; Kim-Hellmuth et al., 2017).

De la même manière l'étude des différences d'expression entre espèces apparentées permet d'identifier des fonctions sur lesquelles les pressions de sélection ont potentiellement été les plus fortes (Karaman et al., 2003 ; Nowick et al., 2009). Par ailleurs, l'étude de l'évolution étant étroitement liée à l'étude du développement, plusieurs études ont utilisé des comparaisons d'expression entre espèces au cours du développement pour identifier les fonctions conservées au cours de l'évolution, et identifier les phases du développement pendant lesquelles la probabilité d'émergence de nouvelles fonctions était la plus élevée (McCarroll et al., 2004 ; Gerstein et al., 2014 ; Cardoso-Moreira et al., 2019). Ainsi, ces études tendent à montrer que les phases les moins contraintes du développement sont les phases précoces et tardives, allant dans le sens de la théorie du sablier (*hourglass model* en anglais, voir figure 1.5) (Prud'homme & Gompel, 2010 ; T Kalinka et al., 2010).

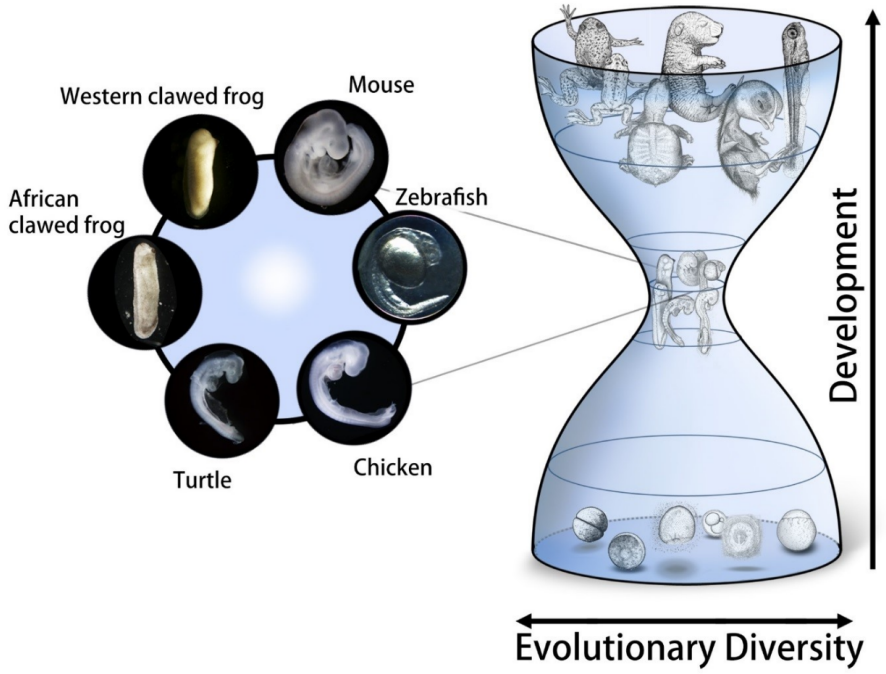

Fig. 1.5 Le modèle du sablier

Au cours du développement, les vertébrés passent par une étape de contrainte élevée au cours de laquelle les embryons d'espèces différentes se ressemblent énormément. Les gènes exprimés au cours de cette période sont en général plus conservés que le reste du génome. Figure tirée de (Irie et al., 2018)

Nous avons vu ici les bénéfices liés à l'étude de la variabilité de l'expression à plusieurs niveaux, ceci nous a permis d'entr'apercevoir ce que de tels travaux pourraient apporter quant à l'évolution et la différenciation de l'expression génique, à l'échelle de l'espèce humaine mais aussi à l'échelle plus large de la comparaison inter-espèces. Par exemple, l'observation d'une réduction significative dans la variabilité de l'expression d'une fonction précise pourrait être le signe d'une adaptation à un nouvel environnement, à l'échelle du transcriptome, rapidement fixée par les processus de sélection naturelle (J. Li et al., 2010).

1.4.3 L'importance de la régulation de l'expression génique dans le cadre de l'évolution

Nous venons de voir l'importance de l'expression des gènes dans l'établissement du phénotype d'un individu au cours de son développement, mais aussi au cours de la vie puisque la modification de l'expression de certains gènes peut être à l'origine du développement de maladies (J. Li et al., 2010 ; Bhatlekar et al., 2014). D'autres exemples soulignant l'importance des mécanismes de régulation de l'expression dans l'établissement d'un phénotype nous viennent des études d'association entre génotype et phénotype, à l'échelle du génome (GWAS, pour *Genome-Wide Association Studies*). En effet, ce type d'étude a permis d'établir que de nombreux phénotypes pouvaient être expliqués, au moins en partie, par des mutations de la séquence d'ADN (Welter et al., 2014 ; Buniello et al., 2019). Néanmoins, une large majorité des variants génétiques ainsi identifiés sont localisés dans des régions non-codantes de l'ADN, et on soupçonne que leur effet sur le phénotype puisse être dû à leur activité régulatrice de l'expression (Schaub et al., 2012 ; Coetzee et al., 2016). De la même manière, il a été démontré que les SNP associés à l'expression de gènes sont significativement enrichis en variants génétiques détectés par GWAS (Schadt et al., 2008 ; Nicolae et al., 2010 ; Zhong et al., 2010 ; M. N. Lee et al., 2014 ; Kim-Hellmuth et al., 2017 ; Piasecka et al., 2018). De par son rôle primordial dans l'établissement des phénotypes, la régulation de l'expression des gènes devrait donc représenter un enjeu crucial dans la lutte pour la survie imaginée par Malthus et théorisée par Darwin (L. Wang et al., 2018).

En effet de nombreuses études semblent confirmer l'importance des mécanismes de régulation de l'expression vis à vis des processus de sélection naturelle. En particulier, il a été démontré de façon récurrente que des changements d'expression génique pouvait être la base moléculaire de phénomènes adaptatifs (King & Wilson, 1975 ; Wray et al., 2003 ; Lopez-Maury et al., 2008 ; Romero et al., 2012 ; Pardo-Diaz et al., 2015). De plus, il existe plusieurs exemples de différences d'expression sur des gènes précis ayant conduit directement à des changements phénotypiques adaptatifs (Abzhanov et al., 2006 ; Chan et al., 2010 ; McBride et al., 2014), et même quelques exemples de divergence transcriptionnelles résultant en une barrière reproductive, suggestifs du potentiel de spéciation des changements d'expression (Haerty & Singh, 2006 ; Kradolfer et al., 2013 ; Thomae et al., 2013 ; Chung et al., 2014 ; Dion-Cote et al., 2014). Enfin, une étude récente a montré l'importance des régions régulatrices et en particulier l'impact de la reconversion d'*enhancers* en promoteurs dans l'histoire évolutive de l'espèce humaine (Carelli et al., 2018). Finalement, le potentiel évolutif des mutations régulatrices de l'expression a été discuté en comparaison avec celui des mutations codantes. En effet, il a été suggéré que l'effet des mutations régulatrices est majoritairement co-dominant, au contraire de celui des mutations codantes qui est plus souvent récessif. En conséquence, l'effet de la sélection naturelle sur des mutations régulatrices sera plus rapide, puisque ne nécessitant pas de dérive génétique pour augmenter en fréquence et atteindre l'état homozygote avant d'être visible par la sélection naturelle (Wray, 2007).

Chapitre 2

Les marques épigénétiques : variabilité et héritabilité au travers de l'exemple de la méthylation de l'ADN

Dans la seconde moitié du XX^{è} siècle et en particulier depuis l'essor des méthodes de séquençage à haut débit, un nombre croissant d'études se sont efforcées d'expliquer les liens existant entre la variance des différents traits phénotypiques et la variance génétique. Cependant pour de nombreux phénotypes, les chercheurs se sont retrouvés dans l'incapacité d'expliquer la variance de ces traits phénotypiques par les seuls facteurs génétiques. En particulier les études sur jumeaux monozygotiques ont permis d'établir que le fait de partager le même génome n'aboutissait pas forcément à un phénotype commun, comme dans le cas de certaines maladies complexes. Bien que l'influence de l'environnement permette d'expliquer une partie de ce phénomène, ces observations mettent en lumière l'impact d'un mécanisme supplémentaire dans la mise en place des phénotypes, mécanisme qui ne repose pas directement sur la séquence de l'ADN mais sensible à l'environnement : c'est l'épigénétique.

2.1 L'épigénétique, architecte de la conformation de l'ADN

2.1.1 Définition de l'épigénétique

L'étymologie du terme épigénétique nous apprend qu'il est formé par l'ajout du préfixe *épi-*, du grec ancien $\varepsilon\pi\iota$ qui signifie «sur», à *génétique*, indiquant donc l'ajout d'une seconde couche d'information à celle préexistante donnée par la séquence de l'ADN. Ce terme, originellement formé par Waddington en 1940 et représenté par les célèbres vallées de Waddington (figure 2.1), désigne au départ la branche de la biologie visant à étudier les mécanismes de causalité entre les gènes et leurs produits qui permettront d'aboutir in fine au phénotype d'un individu (Waddington, 1940, 1957). La définition de l'épigénétique a beaucoup évolué depuis, en 1990 pour Robin Holliday c'est «l'étude des mécanismes de régulation spatiaux-temporels de l'activité des gènes au cours du développement des organismes complexes», une définition qui reprend la vision développementale de Waddington (Holliday, 1990). En 1996, Arthur Riggs et ses collègues proposent une nouvelle définition : « l'étude des changements héritables, au cours de la méiose et/ou de la mitose, de la fonction des gènes qui ne peuvent être expliqués par un changement de la séquence

de l'ADN» (Russo et al., 1996). Plus récemment, Adrian Bird a tenté de se restreindre au contrôle de la conformation de la chromatine en définissant l'épigénétique comme «l'adaptation structurelle des régions chromosomiques visant à sauvegarder, signaler ou perpétuer des modification de l'état d'activité [des gènes]» (A. Bird, 2007). Bien que cette définition s'affranchisse des contraintes imposées par le pré-requis de l'héritabilité qui peut se révéler être problématique, elle pose problème en n'incluant pas un certain nombre de marques épigénétiques telles que l'activité des ARN non-codants ou encore la méthylation de l'ADN des promoteurs que nous décrirons plus en détail ci-dessous. Une vision plus récente nous vient de Michael Skinner qui définit l'épigénétique comme «les facteurs moléculaires et processus prenant place autour de l'ADN, stables au cours de la mitose et qui régulent l'activité du génome indépendemment de la séquence de l'ADN» (Skinner et al., 2010). Finalement, puisque que parue dans un journal plus généraliste et donc destinée à un public plus large que la communauté scientifique mentionnons la définition donnée par l'édition en ligne du journal Le Monde : «l'ensemble des changements d'activité des gènes qui sont transmis au fil des divisions cellulaires ou au fil des générations sans faire appel à des mutations de l'ADN» (Rosier, 2012).

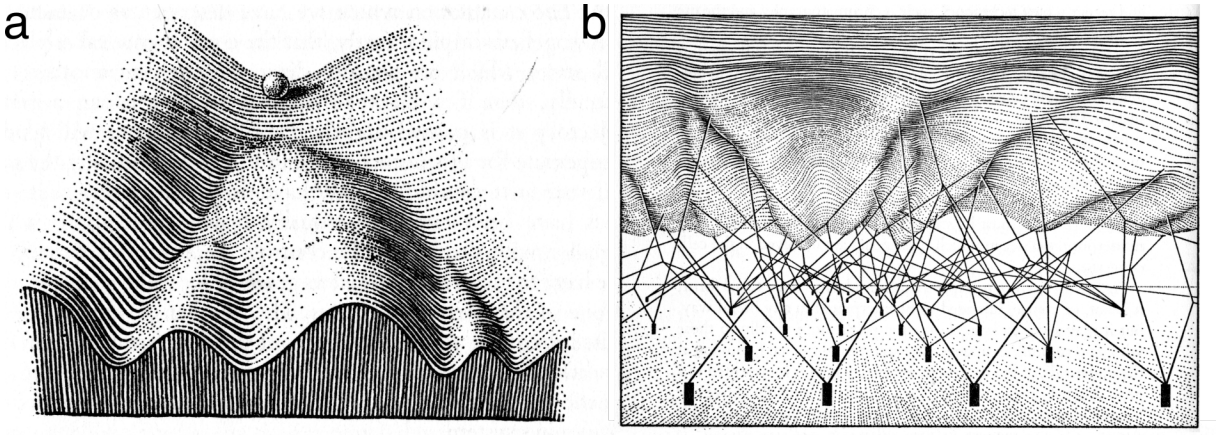

FIGURE 4
Part of an Epigenetic Landscape. The path followed by the ball, as it rolls down towards the spectator, corresponds to the developmental history of a particular part of the egg. There is first an alternative, towards the right or the left. Along the former path, a second alternative is offered; along the path to the left, the main channel continues leftwards, but there is an alternative path which, however, can only be reached over a threshold.

FIGURE 5
The complex system of interactions underlying the epigenetic landscape. The pegs in the ground represent genes; the strings leading from them the chemical tendencies which the genes produce. The modelling of the epigenetic landscape, which slopes down from above one's head towards the distance, is controlled by the pull of these numerous guy-ropes which are ultimately anchored to the genes.

Fig. 2.1 Paysages épigénétiques de Waddington

a) Le paysage épigénétique de Waddington décrit les différentes voies développementales que peut emprunter une cellule, représenté par une bille suivant des vallées. b) Le système complexe d'interaction à l'origine du paysage épigénétique décrit en a). Les gènes sont représentés par les petits tirets noirs. Figures et légendes tirées de la publication originale (Waddington, 1957).

L'épigénome, l'état épigénétique de la cellule, contribue à déterminer l'ensemble des profils d'expression au sein des différents types cellulaires qui constituent le corps humain, et ce alors que chaque cellule possède le même génome. Voyons maintenant en détail les différentes marques composant l'épigénome avant de nous concentrer sur l'étude d'une marque en particulier, la méthylation de l'ADN.

2.1.2 Les différentes marques épigénétiques

Comme nous venons de le voir, l'état de la chromatine dans les cellules est intimement lié aux processus d'expression, et de ce fait aux différentes marques épigénétiques (comme précisé par la définition d'Adrian Bird). En particulier, deux marqueurs sont principalement impliqués dans l'établissement des profils d'euchromatine et d'hétérochromatine : la méthylation de l'ADN que nous aborderons dans la partie suivante, et les modifications des queues N-terminales des histones (Figure 2.2). Comme nous l'avons précédemment abordé les histones sont des protéines essentielles à l'établissement du coeur protéique des nucléosomes, toutefois nous ne nous sommes pas attardés sur le rôle des chaînes N-terminales de ces protéines qui sont sensibles à tout un registre de modifications. On peut en distinguer une dizaine, toutes possédant des rôles différents dans la régulation de la transcription, la réplication, la réparation et la condensation de l'ADN (Kouzarides, 2007 ; Bannister & Kouzarides, 2011). Parmi les plus importantes, ou tout du moins les plus étudiées, citons l'*acétylation* des lysines (correspondant à l'ajout d'un groupe acétyle –CO-CH_3) que l'on retrouve en général dans l'euchromatine et associée à des gènes activement transcrits. En fonction de son emplacement le long des gènes et de la position des résidus modifiés le long des queues d'histones, la *méthylation* des lysines (ajout d'un groupe méthyle -CH_3) et des arginines est associée pour sa part soit à une activation de l'expression, soit à une répression de l'expression et à l'hétérochromatine. Certaines modifications impliquent l'addition de polypeptides bien plus longs, comme c'est le cas pour l'*ubiquitinylation* et la *sumoylation* des lysines qui sont aussi impliquées dans la régulation de l'activité transcriptionnelle. De manière générale, il semble admis que le mode d'action de ces modifications des queues d'histones repose sur le recrutement d'enzymes chargées de remodeler la structure de la chromatine de manière locale, en repositionnant les nucléosomes le long du génome. Ainsi, la co-localisation de certaines modifications des queues d'histones avec la méthylation de l'ADN semble constituer un codé épigénétique déterminant l'état de la chromatine (Strahl & Allis, 2000 ; Szulwach & Jin, 2014 ; Roadmap Epigenomics et al., 2015). Enfin, bien que les mécanismes ne soient pas clairement identifiés il a été suggéré que ces marques sont transmises lors des divisions cellulaires (Hansen et al., 2008 ; Budhavarapu et al., 2013 ; Reveron-Gomez et al., 2018).

Comme nous l'avons précédemment mentionné, il existe un autre type de marque épigénétique, cible d'un intérêt grandissant : les ARN non-codants (ARNnc). Ces ARN sont impliqués dans des fonctions aussi diverses que la régulation de l'expression au cours du développement, l'empreinte parentale et le remodelage de la chromatine (Esteller, 2011 ; Cao, 2014 ; Taylor et al., 2015 ; Uszczynska-Ratajczak et al., 2018). Il existe une grande variété des ARNnc par leur taille, leurs mécanismes d'action ou leurs fonctions, nous nous intéresserons à quelques exemples emblématiques, l'un des plus étudiés étant sûrement Xist. Cet ARNnc est impliqué dans l'extinction de l'un des deux chromosomes X chez la femme à une étape du développement embryonnaire. Cette répression étant initialement aléatoire, puis transmise fidèlement au cours des divisions cellulaires il en découle que toutes les cellules d'une même lignée partageront la même version du chromosome X, mais qu'on pourra observer des différences d'une lignée cellulaire à l'autre (Plath et al., 2002). Une démonstration très visuelle de ce phénomène est trouvée dans le pelage des chats calico ou écailles de tortue, qui trouve son explication dans le positionnement d'un des gènes responsables de la couleur du pelage sur le chromosome X (Kalantry, 2011). Parmi les petits ARNnc impliqués dans la régulation de l'expression, distinguons les ARNpi – des ARN interagissant avec les protéines Piwi – des microARN qui sont des petits ARN d'environ 22 pb dont la classification comme acteurs épigénétiques est encore

matière à débat. Les ARNpi sont impliqués dans la méthylation *de novo* de l'ADN, en particulier au cours des premiers stades du développement, ainsi que dans la répression des éléments transposables (Malone & Hannon, 2009 ; Juliano et al., 2011 ; Ozata et al., 2019). Ces séquences d'ADN, extrêmement fréquentes dans notre génome, sont parfois capables de se déplacer de manière autonome dans le génome, avec le plus souvent des conséquences délétères. Les microARN ont eux aussi un rôle régulateur de l'expression génique, toutefois leur action se situe entre l'étape de transcription et de traduction, par séquestration et/ou clivage des ARN messagers (Bartel, 2009 ; Catalanotto et al., 2016 ; O'Brien et al., 2018). C'est parce qu'ils interviennent après la transcription que certains ne les considèrent pas comme des acteurs épigénétiques, toutefois il ne fait aucun doute que l'activité des microARN est intimement reliée à celle des différentes marques épigénétiques (J. C. Chuang & Jones, 2007 ; Sato et al., 2011).

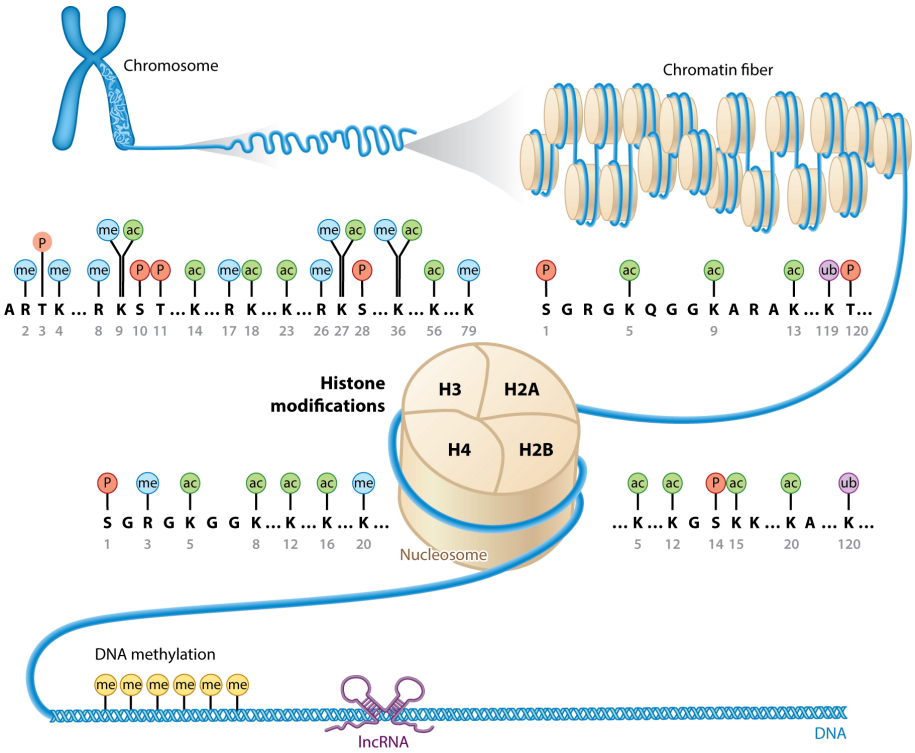

Fig. 2.2 Les différents acteurs épigénétiques
Figure tirée de (Z. Chen et al., 2017)

Enfin, un certain nombre de protéines sont impliquées dans le maintien des patrons d'expression spatio-temporels mis en place au cours du développement. Ces protéines, à savoir Polycomb et Trithorax, forment des complexes qui jouent un rôle primordial dans la mise en place et le maintien de programmes d'expression induits au cours du développement, mais aussi par l'environnement. Ainsi, il a été montré que PRC1 et PRC2, les deux principaux complexes où est retrouvé Polycomb, sont responsables de la méthylation de la lysine 27 de l'histone H3, et de l'ubiquitinylation de la lysine 119 de l'histone H2 (Schuettengruber et al., 2017).

2.2 La méthylation de l'ADN

Nous avons vu les principales marques épigénétiques et leurs différentes fonctions, il est temps maintenant de s'intéresser à la méthylation de l'ADN, la marque épigénétique la plus étudiée de par son accessibilité et sa stabilité ainsi que son caractère informatif quant à l'activité transcriptionnelle des gènes. Pour toutes ces raisons, c'est la marque épigénétique qui a été utilisée au cours de cette thèse comme indicateur de l'état épigénétique.

2.2.1 Dans les différents domaines du vivant

La méthylation de l'ADN est l'une des marques épigénétiques les plus anciennes puisqu'on la retrouve dans les 3 grandes branches du vivant. Consistant chez les eucaryotes en l'ajout d'un groupe méthyle aux cytosines, en position C5, elle s'effectue en règle générale dans le contexte de dinucléotides cytosine guanine (symbolisés par CpG, le p indiquant le pont phosphate entre la cytosine et la guanine) (Moore et al., 2013). Toutefois, la méthylation de l'ADN peut se retrouver dans des contextes différents en fonction de l'espèce considérée. Ainsi, chez les plantes on retrouve une méthylation des cytosines dans le cadre de trinucléotides CHG, ou même de manière asymétrique dans le cadre de CHH (H symbolisant soit un C, T ou A) (T. F. Lee et al., 2010). Chez les mammifères, et l'espèce humaine en particulier on retrouve une immense majorité des cytosines méthylées dans le contexte de CpG, toutefois une faible portion affecte des dinucléotides CpA et CpT, pour une méthylation ici aussi asymétrique (Jin et al., 2011 ; Moore et al., 2013). Ces proportions sont dépendantes du type cellulaire puisque si on n'observe que 2% de méthylation asymétrique dans les cellules somatiques humaines, cette proportion augmente jusqu'à 25% dans les cellules souches embryonnaires (Jin et al., 2011). Cette distinction est importante puisque la symétrie de la méthylation sur les deux brins d'ADN est un mécanisme très important pour sa préservation au cours des divisions cellulaires successives, un point sur lequel nous reviendrons dans les parties suivantes.

Enfin la méthylation de l'ADN est aussi retrouvée chez les procaryotes, à la fois chez les bactéries et les archées (Casadesus & Low, 2006 ; Blow et al., 2016). Toutefois dans ces domaines du vivant, la méthylation est bien plus répandue et peut affecter les adénines (6-méthyladénine, 6mA), et les cytosines en position 4 et 5 (4mC, et 5mC) (Korlach & Turner, 2012 ; Marinus & Lobner-Olesen, 2014). Notons ici que la méthylation des adénines a aussi été observée récemment dans des cellules souches embryonnaires de mammifères, mais leur rôle reste peu étudié en dehors des procaryotes (T. P. Wu et al., 2016). Par ailleurs, des différences majeures existent dans la fonction de la méthylation au sein de ces organismes. En effet, si son rôle est plus lié à la structure de la chromatine et à la régulation de l'expression chez les mammifères, la méthylation de l'ADN revêt en outre un rôle primordial d'identité et de reconnaissance du "non-soi" chez les bactéries. Au travers d'un système dit de restriction-modification, les bactéries peuvent en effet discerner leur ADN de celui des virus les infectant, via la non-méthylation des ADN étrangers (Arber, 1974 ; Casadesus & Low, 2006). En outre, la méthylation de l'ADN est aussi impliquée dans des processus de réparation de l'ADN en cas de coupure simple brin et dans l'initiation de la réplication (Lobner-Olesen et al., 2005).

Répandue dans l'ensemble des domaines du vivant, la méthylation de l'ADN est une marque épigénétique primordiale au bon fonctionnement des cellules comme nous avons pu l'entrevoir. L'ensemble de ses caractéristiques, de ses mécanismes de maintien et d'éta-

blissement, et ses fonctions est trop vaste pour être détaillé ici, nous nous concentrerons donc maintenant sur l'exemple de l'espèce humaine.

2.2.2 Genèse et maintien des profils de méthylation chez l'Homme

Chez l'Homme, l'ajout du groupe méthyle sur une cytosine est la conséquence de plusieurs mécanismes impliquant des enzymes de la famille DNMT (pour « DNA-methyl transferases » en anglais) (Moore et al., 2013). La première enzyme de cette famille, DNMT1, est impliquée dans le maintien des profils de méthylation au cours de la réplication de l'ADN précédant la division cellulaire. Ce maintien est assuré par la méthylation du brin néo-synthétisé par DNMT1, en suivant par symétrie le patron de méthylation du brin parental. On voit apparaître ici l'importance de la symétrie dans le maintien des profils de méthylation que nous avions précédemment évoqué, la méthylation asymétrique des CpA ou CpT étant ainsi graduellement diluée au cours des divisions cellulaires dans les cellules somatiques. La différence de fréquence de la méthylation asymétrique entre cellules somatiques et cellules embryonnaires s'explique potentiellement par l'activité particulièrement élevée de DNMT3A et DNMT3B, en interaction avec DNMT3L, dans les cellules germinales et les cellules souches embryonnaires (Okano et al., 1999). En effet ces enzymes sont impliquées dans la méthylation *de novo* de l'ADN, et leur mécanisme ne repose pas sur la symétrie entre brins, menant donc à un plus fort taux de méthylation hors du contexte habituel CpG dans ces cellules (Lister et al., 2009).

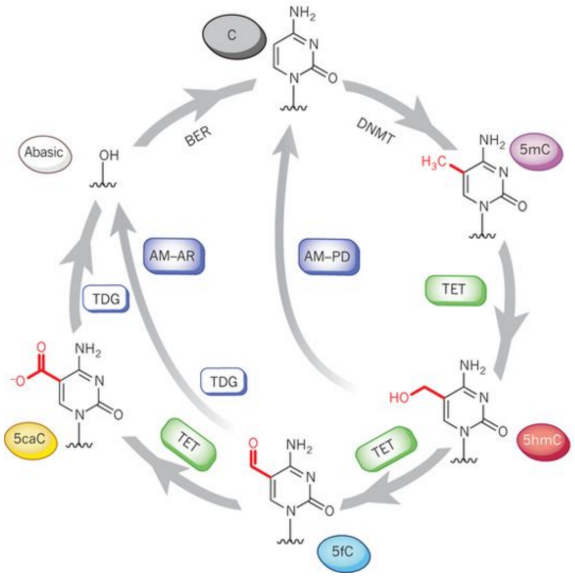

Fig. 2.3 Les différentes voies de déméthylation de l'ADN
On distingue la voie passive par dilution des 5-hmC (AM-PD), de la voie active impliquant l'enzyme TDG (AM-AR). Figure tirée de (Kohli & Zhang, 2013).

La déméthylation active de l'ADN, qui joue un rôle primordial lors des phases précoces du développement embryonnaire, implique l'activité d'un certain nombre d'enzymes et deux mécanismes ont été à ce jour proposés (Feng et al., 2010). La première étape, commune aux deux voies, consiste en l'hydroxylation des 5-mC en 5-hydroxyméthylcytosines (5-hmC) par l'enzyme TET (ten-eleven translocation) (Tahiliani et al., 2009). À partir de 5-hmC, le processus de déméthylation peut se poursuivre de manière passive, par dilution

due à la plus faible affinité de DNMT1 pour les 5-hmC relativement aux 5-mC (Valinluck et al., 2004 ; Hashimoto et al., 2012). L'autre mécanisme, actif, implique le système de réparation par excision de base et nécessite une ou deux oxydations supplémentaires des 5-hmC en 5-formylcytosine (5-fC), puis potentiellement en 5-carboxylcytosine (5-caC), toujours catalysée par l'enzyme TET (He et al., 2011). Ces deux types de cytosine modifiée, 5-fc et 5-caC, sont ensuite reconnus par l'enzyme TDG (thymine DNA glycolase) qui va conduire à leur excision puis leur remplacement par une cytosine non-méthylée (H. Wu & Zhang, 2014) (Figure 2.3).

Ces processus de méthylation et déméthylation sont à l'origine des profils de méthylation au sein des différentes cellules de notre organisme. Ainsi si certaines régions sont réputées peu-variables, telles que les îlots CpG, des régions à fort taux de sites CpG, généralement localisées à proximité des promoteurs des gènes et caractérisées par un faible niveau de méthylation, d'autres régions sont plus variables (A. Bird et al., 1985 ; A. P. Bird, 1986). En effet, bien qu'il soit estimé qu'entre 70 et 80% des sites CpG de notre génome sont méthylés, il existe une certaine variabilité entre types cellulaires, au cours de la vie et entre individus, variabilité qui est reliée à l'activité cellulaire comme nous allons le voir (Smith & Meissner, 2013).

2.3 La variabilité des profils de méthylation de l'ADN chez l'Homme, à plusieurs échelles

L'étude de l'épigénétique, et de la méthylation de l'ADN en particulier, a grandement bénéficié des avancées technologiques en matière de séquençage de l'ADN. En effet, avec l'évolution et la popularisation de méthodes de séquençage à faible coût, se sont développées plusieurs méthodes d'acquisition des profils de méthylation. Ces différentes méthodes reposent toutes sur une transformation chimique de l'ADN sous l'effet du bisulphite. L'action de cette molécule sur l'ADN a pour conséquence la transformation des cytosines en uraciles, une base azotée proche de la thymine et reconnue comme telle par les méthodes de séquençage du génome. Toutefois cette action du bisulphite est bloquée par la méthylation des cytosines, ainsi les 5-mC et 5-hmC ne sont pas transformées en uracile, et seront reconnues comme des cytosines au cours du séquençage (Patterson et al., 2011). Suite à cette transformation de l'ADN, plusieurs options existent : le séquençage total a l'avantage d'être la méthode la plus exhaustive dans l'établissement des profils de méthylation, mais présente quelques inconvénients. En premier lieu, il ne permet pas de distinguer les 5-mC des 5hmC, ce qui peut néanmoins être résolu en réalisant une étape supplémentaire d'immuno-précipitation en amont du séquençage (Thu et al., 2009). Le désavantage principal du séquençage total provient du coût de cette technique, en effet bien qu'exhaustif, il sera généralement impossible à mettre en place dans le cadre d'études visant à comparer les profils de méthylation d'un grand nombre d'échantillons. Toutefois, il existe une méthode alternative qui implique l'utilisation de puces, dont la plus complète à ce jour est l'Infinium MethylationEPIC Kit d'Illumina, qui capture l'état de méthylation de plus de 850,000 CpG le long du génome, enrichis en îlots CpGs, promoteurs et *enhancers* (Moran et al., 2016).

Le développement de ces méthodes ont permis en premier lieu l'établissement du profil moyen de la méthylation de l'ADN chez l'espèce humaine : pour les gènes fortement exprimés, on observe un faible niveau de méthylation au niveau des promoteurs, tandis que le corps du gène est fortement méthylé (Smith & Meissner, 2013). Les éléments

transposables, les télomères et les centromères sont eux aussi fortement méthylés, tandis qu'il existe un faible nombre de régions pour lesquelles un allèle seulement est méthylé, on parle alors d'hémi-méthylation, c'est le cas des régions à empreinte parentale (E. Li et al., 1993 ; Jones, 2012). Cependant, dans un second temps des écarts à ce profil moyen ont pu être observés, et plusieurs exemples de variabilité des profils de méthylation ont été décrits, à plusieurs échelles.

2.3.1 Au cours de la vie

Ainsi, au sein d'un individu et pour un type cellulaire donné, on peut observer des variations des profils de méthylation au cours de sa vie. Cette variation peut être la conséquence d'un processus normal de vieillissement, on a par exemple montré une hypométhylation globale du génome, opposée à une hyperméthylation des îlots CpG avec l'âge (Bjornsson et al., 2008 ; Bollati et al., 2009 ; Christensen et al., 2009). Cette relation de la méthylation de l'ADN au vieillissement a une application très concrète : en effet, un certain nombre de sites CpG peuvent être utilisés pour déterminer de manière très précise l'âge d'un individu. L'identification de sites dont le niveau de méthylation est lié à l'âge a donné lieu a de nombreuses études, l'objectif étant de capturer le plus petit nombre de sites CpG possible pour construire une horloge épigénétique (*epigenetic clock*) (Hannum et al., 2013 ; Horvath, 2013 ; Weidner et al., 2014 ; M. E. Levine et al., 2018). De plus, quelques études chez la souris ont suggéré que la variabilité des profils de méthylation avec l'âge pourrait jouer directement un rôle dans les processus de vieillissement, même si plus d'analyses sont nécessaires afin de confirmer ces résultats (Unnikrishnan et al., 2019).

Enfin, si la modification de la méthylation de l'ADN au cours de la vie est un processus normal, une autre source de variabilité au cours de la vie est le résultat des pressions environnementales. En particulier il a été démontré que les interactions avec des pathogènes peuvent modifier le profil de méthylation, de telles modifications jouant le rôle de mémoire des infections passées dans les cellules du système immunitaire (Pacis et al., 2015). Par ailleurs, des évènements tels que le développement de cancers, ou l'obésité peuvent provoquer une accélération du processus de vieillissement des patrons de méthylation dans les cellules concernées (Kulis & Esteller, 2010 ; Horvath, 2013 ; Horvath et al., 2014).

2.3.2 Entre types cellulaires

S'il existe un grand nombre de différences liées au vieillissement et à l'histoire de vie des cellules, au sein d'un individu la source principale de variabilité des profils de méthylation est à chercher au niveau des différences entre types cellulaires. En effet, une étude comparant la méthylation de l'ADN entre chimpanzés et humains a montré que les différences de profils de méthylation entre deux tissus au sein de la même espèce sont plus importantes que les différences observables entre les deux espèces pour un tissu donné (Pai et al., 2011). Les raisons d'une telle variabilité entre types cellulaires est à chercher dans leur très forte spécialisation au cours du développement, qui résulte de la répression stable de certains promoteurs pendant la différenciation cellulaire (Suzuki & Bird, 2008 ; Lister et al., 2009 ; Smith & Meissner, 2013). Ainsi de nombreux réseaux de gènes sont verrouillés dans un état actif ou inactif en fonction des patrons de méthylation acquis au cours des

différentes étapes du développement embryonnaire.

Il existe de nombreuses applications à cette variabilité de la méthylation de l'ADN entre types cellulaires. La première vient du profil atypique des cellules cancéreuses que nous avons précédemment mentionné. En effet, si on observe une accélération du vieillissement dans les cellules cancéreuses, la comparaison avec les cellules saines apporte de nouvelles pistes pour améliorer la détection et l'établissement de pronostics (Qureshi et al., 2010 ; Barrow & Michels, 2014 ; Salta et al., 2018). Par ailleurs, ces études comparatives ont permis l'identification de nouvelles cibles thérapeutiques, basée sur l'utilisation de substances hypo-méthylantes (Issa, 2007 ; Da Costa et al., 2017). Un second exemple d'application provient du développement de méthodes prédictives des proportions en type cellulaire. Ces méthodes bio-informatiques utilisent les profils de méthylation acquis dans des tissus hétérogènes, tels que le sang, pour prédire les proportions respectives des différents types cellulaires (Houseman et al., 2012 ; Koestler et al., 2013 ; Titus et al., 2017).

2.3.3 Entre individus et populations

L'épigénétique en général, et la méthylation de l'ADN en particulier, est à l'instar de l'expression des gènes, un phénotype moléculaire qui représente une porte d'entrée aux mécanismes gouvernant le fonctionnement des cellules. De ce fait, un grand nombre d'études se sont intéressées à l'identification de différences de méthylome entre individus. Cette recherche de sites différentiellement méthylés (DMS, *differentially methylated sites*), permet d'entr'apercevoir les causes et/ou les mécanismes sous-jacents à la grande diversité phénotypique de notre espèce. Ce type d'étude peut être mené au sein d'une population homogène, entre hommes et femmes par exemple ou entre personnes saines et malades, mais aussi entre populations. La première approche sera détaillée dans le prochain chapitre et correspond aux études dites d'EWAS, la seconde est en générale menée afin d'améliorer notre compréhension des phénomènes nous ayant permis, et nous permettant de nous adapter à des environnements variés. Toutefois, avant de nous intéresser aux différences de méthylation d'ADN existant entre population, notons qu'il existe bien une certaine variabilité entre individus, même si la détermination des causes de cette variabilité est une tâche qui peut s'avérer ardue (Bock et al., 2008 ; Garg et al., 2018). De plus, la détermination de la gamme des variations de la méthylation de l'ADN au sein d'une population saine est primordiale afin de détecter avec confiance des profils de méthylation aberrants chez des individus malades (Bock et al., 2008). Ce type d'étude permet d'établir des clusters de sites CpG hyper-variables entre individus, qui semblent spécifiques au tissu concerné et réactifs à des modifications de l'environnement, ce qui met potentiellement en lumière des régions du génome impliquées dans l'adaptation à un environnement changeant (Garg et al., 2018).

D'un autre côté, certaines études se sont intéressées aux différences de méthylation qui pouvaient exister entre populations humaines. De plus, les méthodes d'enrichissement des régions du génome différentiellement méthylées en fonctions cellulaires peuvent contribuer à affiner l'identification des régions du génome ayant subi différentes pressions sélectives, relativement aux différences d'environnement des populations étudiées. La combinaison de ces deux approches permet tout du moins d'identifier les fonctions cellulaires ayant le plus divergées entre populations humaines (Heyn et al., 2013). Ainsi, il a été montré que des différences majeures dans la méthylation du sang pouvaient être identifiées entre des populations européennes et africaines, avec entre 13 et 21% de sites testés identifiés

comme étant significativement différentiellement méthylés avec un taux de faux positifs inférieur à 1% (Fraser et al., 2012 ; Moen et al., 2013). En outre, la comparaison de 5 populations permettant une large couverture de l'histoire des migrations de notre espèce a mis en évidence le rôle des différences génétiques dans les variations de profils de méthylation de l'ADN entre populations (Carja et al., 2017). Finalement, une autre étude digne d'être mentionnée est le résultat du travail de Maud Fagny, qui a comparé les profils de méthylation dans le sang entre des populations de chasseur-cueilleurs (RHG) vivant dans la forêt équatoriale en Afrique Centrale, avec des populations d'agriculteurs Bantus (AGR) vivant soit dans la forêt, soit dans des zones urbaines. Ces deux populations s'étant séparées il y a environ 60,000 ans, la conception de ce projet a permis de comparer le profil de méthylation de populations présentant des différences historiques mais partageant le même environnement actuel (RHG vs AGR-forestiers), mais aussi de populations ayant une même histoire évolutive mais différant dans leur environnement actuel (AGR-urbains vs AGR-forestiers). Ceci a permis aux auteurs de cette étude d'identifier plus ou moins le même nombre de différences historiques que de différences récentes, mais affectant des fonctions biologiques profondément différentes. Ainsi, si les différences de méthylation historiques peuvent être reliées à des fonctions biologiques impliquées dans des traits phénotypiques par ailleurs différents entre les deux populations, les différences de méthylation récentes étaient elles enrichies en fonctions immunitaires (Fagny et al., 2015).

2.4 L'origine de la variabilité de la méthylation de l'ADN

Dans l'ensemble des exemples que nous venons de présenter, une partie des différences était attribuable à des différences en fréquence de certains variants génétiques sous-jacents, tandis qu'une fraction non négligeable pouvait être expliquee par l'impact de facteurs environnementaux.

2.4.1 Les facteurs génétiques

Un premier exemple de l'impact des facteurs génétiques sur les profils de méthylation est la conséquence d'études comparant le méthylome de jumeaux monozygotes et dizygotes. Le fait que les profils de méthylation des jumeaux monozygotes étant plus corrélés que ceux des jumeaux dizygotes suggère que des facteurs génétiques ont une influence sur le méthylome (Ollikainen et al., 2010 ; Schneider et al., 2010 ; Bell & Spector, 2012 ; Hannon et al., 2018). Par ailleurs, à l'instar des méthodes de génomique et de transcriptomique tentant d'identifier les facteurs responsables des profils d'expression, des méthodes consistant à associer les profils de méthylation avec le génotype ont vu le jour. Ces méthodes, visant à l'identification de meQTL (pour *methylation Quantitative Trait Loci*, en analogie avec eQTL) dans diverses tissus et populations peuvent se conduire de manière locale, lorsque la distance entre le site CpG et le SNP n'excède pas 1Mb (on parle de *cis*-meQTL) ou à l'échelle du génome, le site CpG et le SNP pouvant même être sur des chromosomes différents (*trans*-meQTL) (Bell et al., 2011 ; Fraser et al., 2012 ; Gutierrez-Arcelus et al., 2013 ; Heyn et al., 2013 ; Moen et al., 2013 ; Wagner et al., 2014 ; Pai et al., 2015). Une majorité des différences de méthylation entre populations s'explique ainsi par une différence de fréquence des meQTL entre les deux populations (Fraser et al.,

2012 ; Moen et al., 2013) (figure 2.4). Toutefois il existe aussi quelques exemples d'effet de meQTL qui sont spécifiques à une population, ce qui suggère l'existence d'interactions entre plusieurs facteurs génétiques ou avec des facteurs environnementaux (effets GxG ou GxE) (Fraser et al., 2012 ; Heyn et al., 2013 ; Moen et al., 2013).

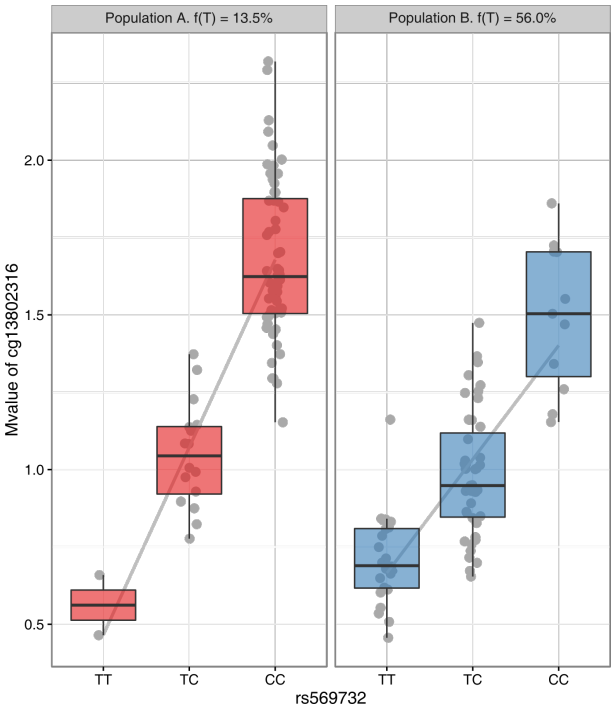

Fig. 2.4 Exemple de meQTL dans deux populations

Le niveau de méthylation du CpG cg13802316 est indiqué en fonction du génotype du variant rs569732, dans deux populations A et B. La fréquence de l'allèle ancestral est indiquée dans le bandeau supérieur pour les deux populations. On peut observer que la différence de fréquence entre les deux populations a pour conséquence un niveau moyen de la méthylation plus élevé dans la population A que dans la population B.

Bien que les mécanismes par lesquels les meQTL impactent le niveau de méthylation des sites CpG auxquels ils sont associés ne soient pas encore déterminés avec certitude, plusieurs études ont suggéré un lien avec l'activité des facteurs de transcription. En effet il a été montré, en particulier pour les *trans*-meQTL, un enrichissement en *cis*-eQTL contrôlant l'expression de facteurs de transcription (Bonder et al., 2017). D'autres études ont démontré que les meQTL étaient enrichis en sites de fixation de facteurs de transcription (Banovich et al., 2014 ; Kaplow et al., 2015). Dans les deux cas, le mécanisme sous-jacent est que les meQTL vont modifier l'activité des facteurs de transcription le long du génome ce qui va conduire à des modifications de la conformation de la chromatine et en conséquence des modifications des profils de méthylation (Zhu et al., 2016 ; Yin et al., 2017).

2.4.2 Les facteurs environnementaux

De la même manière que pour déterminer l'impact des facteurs génétiques, l'étude des profils de méthylation des jumeaux monozygotes et dizygotes a grandement contribué à comprendre l'impact des facteurs environnementaux sur la méthylation de l'ADN (Bell

& Spector, 2012 ; Schneider et al., 2010 ; Ollikainen et al., 2010 ; Hannon et al., 2018). Toutefois les mêmes limites que pour distinguer l'impact des facteurs environnementaux sur les profils d'expression s'appliquent ici. En particulier, il est difficile de distinguer les variations des profils de méthylation dues à l'influence de l'environnement de celles dues à des variations stochastiques ou contrôlées par la génétique.

Malgré ces limites, on sait aujourd'hui que de nombreux facteurs environnementaux ont la capacité d'influer notre méthylome, tels que l'exposition au soleil, le tabagisme, ou l'infection par un certain nombre de pathogènes (Gronniger et al., 2010 ; K. W. Lee & Pausova, 2013 ; Pacis et al., 2015 ; Vandiver et al., 2015 ; Patin et al., 2019). Ces modifications peuvent être liées à des conséquences phénotypiques, comme dans le cas de l'exposition au soleil où les changements de méthylation sont semblables à ceux observés dans le cas de cancer de la peau, ce qui suggère que le risque accru de cancer dû à l'exposition au soleil pourrait impliquer des modifications de notre méthylome (Vandiver et al., 2015). Toutefois, si notre méthylome est variable, dans certains cas les changements induits par les facteurs environnementaux sont partiellement réversibles comme c'est le cas pour le tabagisme. En effet, si fumer induit des changements de méthylation de l'ADN, après un arrêt total du tabagisme on observe une rémission de ces changements au cours du temps chez l'adulte (K. W. Lee & Pausova, 2013 ; Tsaprouni et al., 2014). Ce phénomène n'est toutefois pas observé dans le cadre de l'exposition du foetus à la nicotine, qui semble provoquer des effets sur le méthylome à long terme (Toledo-Rodriguez et al., 2010 ; K. W. Lee & Pausova, 2013).

De manière générale, l'effet de l'environnement sur les profils de méthylation semble être plus important au cours du développement embryonnaire et post-embryonnaire, que durant la vie adulte. En effet la nutrition, l'exposition à des composés toxiques, le niveau socio-économique des parents et plus généralement le stress maternel durant la grossesse sont autant d'exemples de facteurs environnementaux qui ont été identifiés comme affectant le programme épigénétique au cours du développement (Gluckman et al., 2009 ; Lam et al., 2012). L'impact de l'environnement sur l'épigénome est bien plus susceptible de se faire sentir durant le développement, qu'il soit embryonnaire, en particulier, ou post-embryonnaire en conséquence aux très nombreuses divisions cellulaires et aux importants remodelages des profils de méthylation ayant lieu durant cette période d'intense différenciation cellulaire (Aguilera et al., 2010). Un autre argument en faveur de la susceptibilité accrue du méthylome à l'environnement au cours du développement vient du fait que l'effet d'un changement de méthylation de l'ADN affectant une cellule indifférenciée aura plus de chance d'être transmis et amplifié aux lignées cellulaires futures, que si ce changement affectait une cellule déjà différenciée.

Enfin, de plus en plus d'études montrent un lien entre modifications du méthylome par l'environnement lors du développement et la petite enfance et variation de l'expression de certains gènes, suggérant que l'épigénome pourrait servir d'interface entre les pressions liées à notre environnement et la façon dont nous utilisons notre patrimoine génétique (Jaenisch & Bird, 2003 ; Mazzio & Soliman, 2012). De manière plus générale, l'accumulation des preuves de l'impact des facteurs génétiques et environnementaux sur la variabilité de la méthylation de l'ADN souligne l'importance de considérer l'interaction entre ces différents acteurs pour comprendre la diversité phénotypique des populations humaines, un point que nous aborderons dans le prochain chapitre (Y. V. Sun, 2014 ; Patin et al., 2019).

2.5 L'héritabilité de la méthylation de l'ADN

Nous venons de discuter les différents facteurs environnementaux affectant les profils de méthylation chez l'Homme, toutefois bien que nous ayons discuté la réversibilité de tels changements, dans de nombreuses situations de telles modifications ont des effets sur le long terme. Ceci implique qu'au cours des divisions cellulaires les patrons de méthylation sont transmis de la cellule parentale aux deux cellules filles, un mécanisme que nous avons touché du doigt précédemment mais que nous allons détailler ici. Enfin, une grande question liée à l'épigénétique de manière générale et à la méthylation de l'ADN en particulier est de savoir s'il existe une transmission trans-générationnelle des modifications de méthylation acquises au cours de la vie.

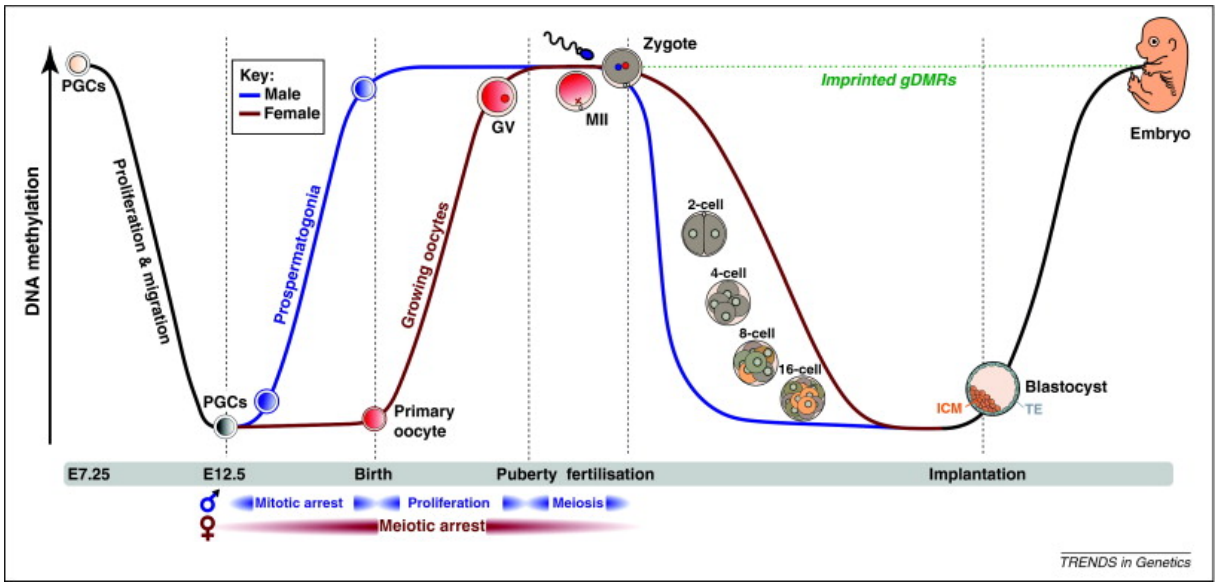

Fig. 2.5 Les deux phénomènes de reprogrammation de la méthylation
La courbe rouge et la courbe bleue représentent le niveau moyen de méthylation du génome maternel et paternel, respectivement, au cours des premières étapes du développement embryonnaire et de la gamétogenèse. Figure tirée de (Smallwood & Kelsey, 2012).

2.5.1 Héritabilité cellulaire

Les profils de méthylation sont transmis très fidèlement au cours de la réplication de l'ADN grâce à l'activité de DNMT1 qui possède une très forte affinité pour les brins d'ADN hémi-méthylés (Hermann et al., 2004 ; Probst et al., 2009). Lors de la réplication de l'ADN, la molécule bi-caténaire de l'ADN va s'ouvrir à la position d'origine, où la réplication va démarrer et deux structures en forme de fourche vont alors se créer permettant la copie de la séquence de l'ADN dans les deux directions. C'est au niveau de cette fourche que DNMT1 est recrutée via une interaction directe avec PCNA qui est impliquée dans la progression de la machinerie de réplication le long de l'ADN et UHRF1 qui se fixe spécifiquement à l'ADN hémi-méthylé via un domaine SRA (*SET and RING associated*) pour recruter DNMT1 (L. S. Chuang et al., 1997 ; Bostick et al., 2007 ; Arita et al., 2008). Enfin DNMT1 est stabilisée via des interactions avec un certain nombre de protéines

associées à la chromatine, telles que LSD1 (*Lysine-Specific Demethylase 1*), essentielle pour réguler le statut de méthylation de DNMT1, mais aussi la méthylation de la lysine 9 de l'histone H3 (H3K9me), qui module la stabilité de DNMT1 durant la phase S de la réplication (J. Wang et al., 2009 ; Rothbart et al., 2012). L'ensemble de ces interactions sont nécessaires à l'activité de maintenance des profils de méthylation par DNMT1, et permettent la grande fidélité de ce processus, en limitant l'activité de DNMT1 au cycle de réplication (Smith & Meissner, 2013).

2.5.2 Héritabilité trans-générationnelle

Si la découverte de variants génétiques associés à des marques de méthylation a permis de proposer un mécanisme de transmission des variations de méthylation causées par des mutations génétiques au cours des générations successives, l'héritabilité des modifications associées à des facteurs environnementaux reste encore à prouver (Heard & Martienssen, 2014). En effet, lors de la gamétogénèse puis de l'embryogénèse, les profils de méthylation sont presque totalement réinitialisés par deux déméthylations complètes du génome (Smallwood & Kelsey, 2012 ; Messerschmidt et al., 2014 ; H. J. Lee et al., 2014) (figure 2.5). Ces étapes étant essentielles à la programmation des cellules embryonnaires en cellules souches, l'héritabilité des profils de méthylation non-associés à des meQTL a été estimée à 0.187 (Feng et al., 2010 ; Cantone & Fisher, 2013 ; McRae et al., 2014). Par ailleurs, en addition aux difficultés inhérentes à cette double reprogrammation des profils de méthylation, il existe une difficulté d'ordre technique pour détecter une héritabilité trans-générationnelle de modification d'origine environnementale de la méthylation de l'ADN. En effet, dans le cas de l'exposition d'une femme enceinte à un facteur environnemental à même de modifier son méthylome, l'impact de ce facteur va en définitive provoquer des changements de la méthylation de l'ADN sur trois générations. En effet le méthylome de la mère va être modifié, mais aussi celui de l'enfant qu'elle porte, puisque directement affecté, et enfin sur la génération suivante, puisque les cellules germinales en cours de différenciation chez l'enfant à naître peuvent elles aussi être impactées (Heard & Martienssen, 2014). Il faut donc, pour observer de véritables phénomènes d'héritage trans-générationel, observer si cet héritage est transmis aux générations F2 (voire F3, dans le cas de l'exposition d'une femme enceinte). Une autre difficulté réside dans le fait que les modifications des profils de méthylation de l'ADN causés par l'environnement a tendance en général à impacter le ou les tissus concernés par le facteur environnemental (la peau dans le cadre de l'exposition au soleil, par exemple). Or, pour que la transmission soit possible il est nécessaire de modifier le méthylome des cellules germinales dans le même temps.

Malgré ces difficultés, il a été démontré chez la souris la transmission d'un caractère lié à la couleur de la queue et des pattes d'un mâle à sa descendance, via un mécanisme impliquant l'activité de microARN (Jablonka & Raz, 2009). Ces petits ARN présentent l'avantage d'être circulant et donc de pouvoir médier l'impact de l'environnement dans les cellules concernées avant d'être transférées aux cellules germinales (Q. Chen et al., 2016). En particulier il a récemment été suggéré que leur activité en *trans* pourrait leur permettre d'échapper au phénomène de reprogrammation épigénétique (Jodar et al., 2013). Aujourd'hui, bien qu'il ait été démontré que des modifications de marques de méthylation causées par une exposition à un environnement pouvaient occasionnellement échapper à la première étape de reprogrammation lors de la gamétogénèse, il est assez généralement

admis que si des phénomènes d'héritabilité trans-générationnelle devaient avoir lieu, ce serait au travers de l'activité de petits ARNnc (Daxinger & Whitelaw, 2012 ; Wei et al., 2014 ; Heard & Martienssen, 2014).

Bien qu'encore matière à débat, cette question de l'héritabilité trans-générationnelle des marques épigénétiques reste passionnante puisqu'elle ramène sur le devant de la scène la question de l'héritabilité des caractères acquis théorisée par Lamarck il y a maintenant plus de deux siècles.

Chapitre 3

Le rôle régulateur de l'expression par la méthylation de l'ADN chez l'Homme

Nous avons décrit dans la première partie comment la diversité phénotypique découle directement de la variabilité de l'expression de l'ADN, et comment la régulation de cette expression par les organismes vivants est l'un des moteurs des processus d'adaptation et de sélection qui ont façonné la biodiversité telle qu'on la connaît à l'heure actuelle. Nous avons mentionné les facteurs environnementaux et génétiques impliqués dans ce contrôle spatio-temporel des profils d'expression, ainsi que les différentes régions génomiques impliquées dans cette régulation. Ceci nous a permis de constater l'importance de la conformation de l'ADN dans la régulation de l'expression génique, et en particulier de son niveau de condensation, qui lui-même est largement contrôlé par les différentes marques épigénétiques du génome. De ce fait il apparaît que les marques épigénétiques, et en particulier la méthylation de l'ADN jouent un rôle, tout du moins indirect, dans les processus de régulation de l'expression génique.

3.1 Les phénotypes associés à la variabilité de la méthylation

3.1.1 Études d'association à l'échelle du génome (EWAS)

Récemment, un nombre croissant d'études se sont concentrées sur la détection d'associations entre variations du méthylome et certains traits phénotypiques. Ces études, à l'échelle du génome, sont nommées EWAS, par analogie aux études de type GWAS que nous avons précédemment décrites (pour *Epigenome-Wide Associations Studies*) (Rakyan et al., 2011 ; Michels et al., 2013 ; Birney et al., 2016). Néanmoins, étant donné la très forte variabilité entre types cellulaires de la méthylation de l'ADN, il existe un certain nombre de limitations à cette méthode. En premier lieu, le choix du type cellulaire ou du tissu doit être fait en accord avec le phénotype étudié (Michels et al., 2013). La seconde limitation découle directement de la première puisque dans le cas où le tissu d'étude est composé de plusieurs types cellulaires différents, il faudra impérativement corriger pour cette hétérogénéité, au risque de détecter de fausses associations. C'est pour ce type d'application que les méthodes de prédiction des proportions des types cellulaires que nous avons décrites ci-dessus prennent tout leur intérêt. Enfin, l'interprétation des résultats d'EWAS n'est pas aussi simple que dans le cadre de GWAS, en partie à cause de la plus grande plasticité des profils de méthylation, relativement à celle des variants génétiques (Birney et al., 2016).

Parmi l'ensemble des EWAS conduites chez l'Homme, notons quelques exemples non-exhaustifs qui donnent une idée de la diversité de traits phénotypiques ayant été associés à la variabilité de la méthylation de l'ADN chez l'Homme. Ainsi la méthylation de l'ADN dans les lymphocytes a été associée à des cas de lupus érythémateux, d'arthride rhumatoïde, de sclérose en plaques ainsi qu'au diabète de type I (Picascia et al., 2015). D'autres études ont montré que la méthylation dans le sang semble être associée à des traits aussi divers que la sensibilité à la douleur, l'indice de masse corporelle ou la schizophrénie (Bell et al., 2014 ; Aberg et al., 2014 ; Wahl et al., 2017). Toutes ces études semblent pointer dans la direction d'un lien fonctionnel entre la méthylation de l'ADN à certains sites CpG et une vaste diversité de traits phénotypiques. Toutefois, comme précédemment précisé, l'interprétation des résultats d'EWAS est un exercice compliqué, et la direction de ce lien (qui du phénotype ou de la méthylation influence l'autre ?) ainsi que la causalité de l'association (est-elle dépendante ou indépendante d'autres facteurs génétiques ou épigénétiques) restent inconnues. De ce fait, il a été proposé une méthodologie pour une "deuxième génération" d'EWAS qui repose sur une meilleure définition des hypothèses initiales, afin de mieux choisir le tissu d'étude et sur un échantillonage dans le temps, afin de pouvoir apporter une réponse au problème de la causalité de l'association, et dont les préceptes sont détaillés par Tuuli Lappalainen et John M. Greally (Lappalainen & Greally, 2017).

3.1.2 Méthylation et réponse immunitaire

Au travers des exemples de résultats d'EWAS que nous venons de détailler, il s'avère que la méthylation de l'ADN joue un rôle particulier dans notre interaction avec les différents pathogènes présents dans notre environnement. Ce lien peut paraître évident tant la méthylation de l'ADN telle que nous l'avons décrite est à l'interface entre le génome et l'environnement. En effet, par sa labilité la méthylation permet aux cellules du système immunitaire de s'adapter rapidement à l'infection par un virus ou une bactérie. En particulier, une étude sur la réponse de cellules dendritiques à une souche virulente de *Mycobacterium tuberculosis* – l'agent bactérien responsable de la tuberculose chez l'Homme – a montré une déméthylation rapide et active de plusieurs milliers de sites CpG suite à l'infection (Pacis et al., 2015). Puisque pérennes au cours du temps, ces modifications suggèrent un rôle de la méthylation de l'ADN dans l'entraînement de l'immunité innée (*trained immunity*). Cette idée, relativement nouvelle, accorde aux cellules de l'immunité innée une mémoire des infections passées permettant une réponse plus efficace aux infections futures et est décrite minutieusement dans cette revue (Netea et al., 2016). Plusieurs études ont montré que des modifications des histones pourraient être responsables de cet entraînement de l'immunité chez l'Homme, quand d'autres ont montré l'implication de la méthylation de l'ADN chez les plantes ou de l'ARN chez les invertébrés (Saeed et al., 2014 ; Lopez Sanchez et al., 2016 ; Castro-Vargas et al., 2017 ; Kaufmann et al., 2018).

Par ailleurs, l'impact de la méthylation de l'ADN sur la différenciation et la fonction des cellules de l'immunité acquise a été largement démontré (P. P. Lee et al., 2001 ; Chappell et al., 2006 ; Scharer et al., 2013 ; Komori et al., 2015). De plus, il existe de nombreuses études suggérant le lien entre dérégulation des profils de méthylation et le développement de cancer, résumées par Marta Kulis et Manel Esteller (Kulis & Esteller, 2010). Enfin, il convient de discuter du rôle de la méthylation de l'ADN dans l'établissement des empreintes parentales : pour un certain nombre de gènes, environ 80 chez

l'Homme, l'allèle paternel ou maternel est constitutivement méthylé et en conséquence seul l'autre allèle est exprimé (D. H. K. Lim & Maher, 2010). Ces gènes sont essentiels au cours du développement de l'embryon et des circuits neuronaux, et une altération de leur profil de méthylation peut conduire à un certain nombre de maladies développementales, telles que les syndrome de Beckwith–Wiedemann, Russell–Silver, Prader–Willi and Angelman (D. H. K. Lim & Maher, 2010). Au vu du rôle déterminant de la méthylation de l'ADN sur la santé, il est logique que de plus en plus d'études s'intéressent aux pistes offertes par cette marque épigénétique pour le développement de nouvelles thérapies, la détection de nouvelles cibles potentielles ou de bio-marqueurs pour améliorer la rapidité et la précision des diagnostics (Issa, 2007 ; D. H. K. Lim & Maher, 2010 ; Suarez-Alvarez et al., 2012 ; X. Yang et al., 2014 ; Jones et al., 2016 ; Saldanha & Tollefsbol, 2018).

3.2 Implication de la méthylation de l'ADN dans la régulation de l'expression des gènes

Une caractéristique intéressante du génome humain est qu'entre 60 et 70% des gènes ont un promoteur associé avec un îlot CpG. Ces courtes séquences d'ADN (entre 500 et 2,000 pb) ont un très fort taux de guanines et de cytosines et sont enrichies en sites CpG, relativement au reste du génome (Deaton & Bird, 2011 ; Maston et al., 2006). La particularité de ces îlots CpG est de rester, en règle générale, non-méthylés, au contraire de la majorité des sites CpG le long du génome. De plus ils sont associés à des gènes dits "de ménage" (*housekeeping genes*), dont la fonction est nécessaire à l'ensemble des types cellulaires et donc activement transcrits (A. P. Bird, 1987). On voit donc qu'il existe une corrélation entre niveau d'expression des gènes et niveau de méthylation des sites CpG.

3.2.1 Preuve expérimentale et modèle canonique

L'impact de la méthylation de l'ADN sur l'expression des gènes est un phénomène connu de longue date : il faut en effet remonter à la fin des années 1970, période à laquelle plusieurs études ont utilisé des enzymes de restriction bactériennes dans différents organismes modèles pour identifier des corrélations entre différences de méthylation et différences d'expression entre tissus (Waalwijk & Flavell, 1978 ; McGhee & Ginder, 1979). Ces études avaient permis de suggérer l'hypothèse que la méthylation puisse jouer un rôle actif dans la régulation de l'activité des gènes, hypothèse qui fut ensuite prouvée expérimentalement via la transfection de gènes méthylés ou déméthylés *in vitro* dans des cellules de souris (Pollack et al., 1980 ; Wigler et al., 1981 ; Stein et al., 1982). Pour compléter ces études historiques qui permirent d'entrevoir le rôle de la méthylation dans la régulation de l'expression, un certain nombre de travaux utilisèrent les propriétés de la 5-azacytidine — un composé proche de la cytosine et qui peut-être incorporé à l'ADN en lieu et place de cette base — et dont les résultats sont résumés dans cette revue (Razin & Cedar, 1991).

Ces premiers pas furent suivis de plusieurs décennies de travaux de recherche qui ont permis d'affiner les hypothèses initialement émises. En particulier le développement de la méthode d'immuno-précipitation des fragments d'ADN méthylés à l'aide d'un anticorps dirigé contre les 5-mC (MedIP, pour *methylated DNA immunoprecipitation*) permit d'établir les premiers profils de méthylation à l'échelle du génome (X. Zhang et al., 2006 ;

Zilberman et al., 2007). Si les études initiales s'étaient concentrées sur la méthylation de l'ADN en aval des gènes, dans la région des promoteurs, de tels profils à l'échelle du génome ont permis de s'intéresser à la méthylation des régions intra-géniques. En conséquence, a émergé un modèle canonique pour le contrôle de l'expression des gènes par la méthylation de l'ADN. Si, en accord avec les premières études, la méthylation des promoteurs est en général associée avec la répression de la transcription, il apparaît au contraire que la méthylation du corps des gènes est corrélée avec un plus fort taux d'expression (Maunakea et al., 2010 ; Jjingo et al., 2012 ; Jones, 2012 ; Wagner et al., 2014 ; X. Yang et al., 2014 ; Gutierrez-Arcelus et al., 2015). Notons toutefois que cette association positive est dépendante du tissu étudié, puisque dans les cellules du cerveau qui ne se divisent pas ou peu la méthylation inter-génique ne semble pas corrélée à une augmentation de l'expression (Aran et al., 2011 ; Guo et al., 2011). Cette dichotomie suggère qu'il existe potentiellement différents modes d'action par lesquels la méthylation de l'ADN pourrait modifier l'activité transcriptionnelle des gènes.

3.2.2 Mécanisme d'action

En effet, bien que de nombreuses zones d'ombres persistent, plusieurs mécanismes ont été proposés par lesquels la méthylation de l'ADN pourrait jouer un rôle dans la régulation de l'expression des gènes. En premier lieu, il a été démontré que l'affinité de certaines classes de facteurs de transcription pour leurs sites de fixation respectifs est dépendante du statut de méthylation des sites CpG à proximité (Yin et al., 2017). Dans cette situation, la méthylation de l'ADN peut être directement responsable d'une augmentation ou d'une répression de l'expression en favorisant ou en inhibant la fixation de facteurs de transcription. Un autre mécanisme, plus indirect repose sur la fixation de protéines possédant des domaines de fixation aux 5-mC, connus sous le nom de MBP (*Methyl-Binding Proteins*). Ces protéines sont considérées comme les décodeurs de l'épigénome, et en particulier de la méthylation de l'ADN : en se fixant à l'ADN méthylé, elles contribuent au recrutement de co-répresseurs qui vont conduire à l'extinction de la transcription (Du et al., 2015). Ces MBP interagissent avec un grand nombre de protéines, certaines permettant de modifier d'autres marques épigénétiques telles que l'acéthylation des histones, et finalement font le lien entre la méthylation de l'ADN et la réorganisation de la chromatine (X. Zhang et al., 2017).

La difficulté inhérente à l'étude des mécanismes de régulation de l'expression par la méthylation de l'ADN réside en partie dans leur pluralité, comme nous l'avons vu avec les rôles opposés que la méthylation peut avoir en fonction de sa localisation génomique. Toutefois, plusieurs études s'accordent à établir un mode d'action commun qui réside sur le rôle d'organisateur du génome que la méthylation de l'ADN partage avec les autres marques épigénétiques. En effet, comme nous l'avons précédemment abordé, il existe un lien fort entre état de condensation de l'ADN et niveau d'expression. Par ailleurs les protéines se fixant à l'ADN méthylé permettent le recrutement de la machinerie enzymatique nécessaire à la modification de la chromatine (X. Zhang et al., 2017). On voit ici émerger le rôle essentiel de la méthylation de l'ADN dans la régulation de l'expression : en modifiant l'accessibilité de l'ADN aux machineries enzymatiques impliquées dans la transcription, la méthylation permet de moduler le niveau d'expression des gènes. Cette idée est supportée par une étude parue en 2009 où les auteurs ont distingué deux classes de gènes, ceux dont l'expression requiert l'action des complexes protéiques SWI/SNF, et ceux dont

l'expression est indépendante de SWI/SNF (Ramirez-Carrozzi et al., 2009). Ce complexe protéique permet le déplacement des nucléosomes en utilisant une source d'énergie, or une correspondance existe entre les gènes indépendants de l'activité de SWI/SNF et ceux possédant un îlot CpG en son promoteur. En conclusion, il semble que l'ADN à proximité des îlots CpG soit accessible de façon intrinsèque, sans nécessiter l'activité de complexes protéiques. En outre, il a été suggéré que la méthylation intra-génique soit responsable d'une augmentation de la transcription par la répression de promoteurs secondaires ou de segments d'ADN répétés, présents au sein du gène (Maunakea et al., 2010). Un mécanisme proposé pour l'activité de cette méthylation du corps des gènes rejoint le mode d'action décrit pour les promoteurs à îlots CpG. En effet, Yang et ses collègues ont proposé que la méthylation repositionne les nucléosomes vers les jonctions entre introns et exons, affectant ainsi la régulation des évènements d'épissage (X. Yang et al., 2014).

3.3 La méthylation : marqueur ou acteur de l'expression génique

Comme nous venons de le décrire, la méthylation de l'ADN a été associée à un grand nombre de traits phénotypiques. Par ailleurs, nous avons détaillé le rôle que la méthylation a dans la régulation de l'expression des gènes. Puisque c'est par son rôle régulateur de l'expression que la méthylation peut *in fine* affecter les divers phénotypes que nous avons mentionnés, de nombreuses études se sont attelées à détecter les sites CpG le long du génome dont le niveau de méthylation contrôlait l'expression des gènes adjacents. La détection de ces sites CpG, connus sous l'acronyme d'eQTM (*expression Quantitative Trait Methylation-Loci*, par analogie avec les eQTL), permet en outre de développer notre compréhension des mécanismes de régulation de la transcription, d'identifier des possibles cibles thérapeutiques.

3.3.1 Développement des études d'association à l'échelle du génome (eQTM)

Les études de détection d'eQTM se sont développées au cours des 10 dernières années, le terme eQTM devant d'ailleurs, à ma connaissance, son origine à une publication de 2013 par Gutierrez-Arcellus et ses collègues (Gutierrez-Arcelus et al., 2013). Ces études tirent profit de jeux de données comprenant des données d'expression et de méthylation pour les mêmes individus et peuvent être conduites localement, en *cis*, ou à l'échelle du génome entier, en *trans*. Quand conduites localement, ces études cherchent à associer par un modèle linéaire pour chaque gène la variabilité inter-individuelle de leur expression avec celle de la méthylation de l'ADN à tous les sites CpG dans une fenêtre maximale de $\pm$1Mb autour du gène étudié. Pour les études menées en *trans*, le même principe s'applique sauf que pour chaque gène, la variabilité inter-individuelle de son expression est associée avec celle de la méthylation de chaque site CpG du génome, qu'ils soient localisés sur le même chromosome ou même sur des chromosomes différents. Un certain nombre de précautions sont toutefois nécessaires, à commencer par l'attention donnée au tissu utilisé. En effet, comme nous l'avons précédemment évoqué la méthylation de l'ADN et l'expression des gènes sont deux phénomènes très spécifiques au type cellulaire dans lequel ils se produisent. De ce fait, il est extrêmement important de bien choisir le tissu

d'étude soit en privilégiant des tissus homogènes ou des types cellulaires purifiés, soit en incluant dans le modèle statistique des termes décrivant la composition cellulaire, dans le cas de tissus hétérogènes comme le sang (Teschendorff & Relton, 2018).

Malgré ces limitations, ce type d'étude a permis de détecter les sites CpG dont le niveau de méthylation était associé avec le niveau d'expression des gènes adjacents dans un nombre croissant de tissus cellulaires (van Eijk et al., 2012 ; Gutierrez-Arcelus et al., 2013 ; Grundberg et al., 2013 ; Bonder et al., 2014 ; Wagner et al., 2014 ; Bonder et al., 2017). Bien que difficilement comparables puisqu'impliquant un nombre différent d'échantillons, et donc variant dans leur puissance à détecter des eQTM, les nombreuses études conduites ont permis de faire émerger des convergences dans l'ensemble des tissus étudiés (Table 3.1). Pour commencer, la majorité des associations détectées correspondent à un contrôle négatif de l'expression par la méthylation, entre 51 et 69% selon le tissu étudié (van Eijk et al., 2012 ; Gutierrez-Arcelus et al., 2013 ; Bonder et al., 2014, 2017). Enfin la localisation des eQTM dans les régions géniques ont permis de confirmer le modèle canonique que nous venons de décrire. En particulier, la majorité des études a détecté un enrichissement des eQTM inhibant l'expression dans les promoteurs des gènes, et de ceux activant l'expression dans le corps des gènes (Gutierrez-Arcelus et al., 2013 ; Wagner et al., 2014 ; Bonder et al., 2017).

Référence	Tissu	Taille d'échantillon	Fenêtre d'étude	FDR	eQTM-gènes
van Eijk et al., 2012	sang total	148	500 kb	5%	452
van Eijk et al., 2012	sang total	148	*trans*	5%	157
Gutierrez-Arcelus et al., 2013	fibroblastes	110	50 kb	10%	596
Gutierrez-Arcelus et al., 2013	LCLs	118	50 kb	10%	3,680
Gutierrez-Arcelus et al., 2013	lymphocytes T	66	50 kb	10%	3,838
Grundberg et al., 2013	tissu adipeux	648	1.5 kb	1%	2,334
Bonder et al., 2014	liver	158	250 kb	5%	1,798
Wagner et al., 2014	fibroblastes	62	250 kb	5%	587
Bonder et al., 2017	sang total	2,101	250 kb	5%	3,842

Table 3.1 Résumé des résultats d'eQTM dans différents tissus
Le nombre de gènes associés à au moins un eQTM est indiqué dans la dernière colonne

3.3.2 Analyse de la causalité des associations entre méthylation et expression génique

Nous avons décrit la corrélation entre la non-méthylation des îlots CpG et l'expression des gènes adjacents. Il a par ailleurs été proposé que le statut de transcription active des gènes possédant un îlot CpG dans leur promoteur contribue à la non-méthylation de l'îlot CpG. En effet, la transcription active des gènes est liée à la méthylation de la lysine 4

de l'histone H3, ainsi qu'à la présence du variant H2A.Z, deux phénomènes qui sont des antagonistes à la fixation des enzymes de la famille DNMT (Ooi et al., 2007 ; Zilberman et al., 2008). On voit ici apparaître une autre limitation de l'étude des eQTM : les associations détectées entre niveau de méthylation d'un site CpG et l'expression d'un gène ne reflète souvent qu'une corrélation et ne donne donc pas d'information sur les relations de cause à effet entre ces deux phénomènes. Par ailleurs, un grand nombre d'études ont montré que les eQTM étaient majoritairement sous contrôle génétique, indiquant une forte concomitance des facteurs génétiques et épigénétiques dans la régulation de l'expression (van Eijk et al., 2012 ; Gutierrez-Arcelus et al., 2013 ; Wagner et al., 2014 ; Bonder et al., 2017). Dans ces situations, de simples études d'association ne permettent pas de déterminer les liens de cause à effet existant entre ces trois variables (Pai et al., 2015 ; Teschendorff & Relton, 2018) (figure 3.1).

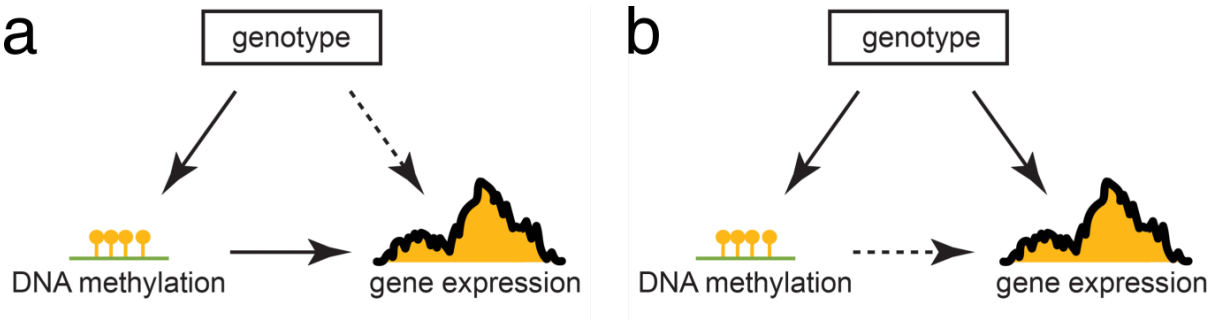

Fig. 3.1 Exemples de relations de cause à effet expliquant une association entre méthylation et expression

Les flèches pleines indiquent des relations causales, tandis que les flèches en pointillées indiquent des corrélations sans relation de cause à effet. a) L'effet de la génétique sur l'expression est médié par la méthylation. b) La génétique contrôle indépendamment à la fois la méthylation et l'expression. Figure adaptée de (Pai et al., 2015).

Afin de surpasser cette limitation, plusieurs méthodes ont été développées, un grand nombre ayant comme point commun d'utiliser les variants génétiques comme une référence fixe puisqu'il est communément admis que des variations de méthylation ou d'expression ont peu de chance d'être responsables d'une mutation de l'ADN. Parmi ces méthodes notons celle reposant sur la corrélation partielle où l'on teste la corrélation entre les résidus de l'expression et de la méthylation après avoir régressé les effets de la génétique (Pai et al., 2015). Les méthodes dites de médiation sont parmi les plus communément utilisées et permettent d'estimer la proportion de l'effet d'une exposition (ici la génétique) sur un résultat (ici l'expression, ou la méthylation) qui est due à une variable intermédiaire, le médiateur (ici la méthylation, ou l'expression) (Teschendorff & Relton, 2018). Enfin, il existe un intérêt croissant porté aux méthodes de randomisation mendélienne (*Mendelian randomization*) qui utilisent le principe des variables instrumentales (*instrumental variables*) et dont les caractéristiques et conditions d'utilisation sont détaillées dans cette revue (Davey Smith & Hemani, 2014).

Toutes ces méthodes permettent d'identifier précisément les sites CpG dont les niveaux de méthylation interfèrent de manière causale avec l'expression. Ceci est d'une très grande importance puisque permet une plus grande précision dans la caractérisation des régions génomiques impliquées dans la régulation de la transcription, et donc *in fine* d'avoir une meilleure compréhension des mécanismes sous-jacents.

Chapitre 4
Objectifs de la thèse

Il est possible de diviser les objectifs de ma thèse en 4 axes de travail principaux, avec en premier lieu, l'étude de la variabilité des profils de méthylation. Comme détaillé dans l'introduction, les profils de méthylation peuvent varier entre populations, d'un sexe à l'autre et au cours de la vie. C'est ces trois axes que j'ai approfondis, en visant à étudier dans un premier temps les différences de méthylation entre individus d'origines européenne ou africaine, puis dans un second temps les différences de méthylation entre hommes et femmes, et entre individus d'âge différent. L'objectif de ce premier axe de travail était d'identifier des sites différentiellement méthylés (DMS), puis les gènes à proximité de ces DMS afin de conduire des analyses d'enrichissement en fonctions biologiques et ainsi avoir une meilleure idée des phénomènes d'adaptation locale qui pourraient avoir eu lieu.

Un second axe de travail était d'identifier les bases génétiques de la méthylation de l'ADN, et en particulier des DMS. La réalisation de ce second objectif a nécessité de cartographier les variants génétiques associés à la variabilité de la méthylation (meQTL), en *cis* et en *trans*. Le recoupement entre DMS et sites CpG sous contrôle génétique devait contribuer à renforcer notre compréhension des différences populationnelles en séparant les différences de méthylation dues à des variations sous-jacentes dans le patrimoine génétique des populations des différences dues à une différence d'environnement. Finalement, l'identification de ces meQTL et l'étude de leur enrichissement en variants précédemment associés à certains traits phénotypiques permettent de mettre en lumière les régions de l'épigénome potentiellement impliquées dans notre adaptation à l'environnement.

Afin d'approfondir ce dernier point, le troisième axe de travail était d'identifier les sites de méthylation directement associés à un phénotype moléculaire : l'expression des gènes. En utilisant des données de séquençage d'ARNm, j'ai identifié les couples gène-CpG pour lesquels le niveau de transcription en ARNm était significativement associé au niveau de méthylation (eQTM). Pour aller plus loin, et observer comment la méthylation de l'ADN pouvait nous permettre de nous adapter à une variation de notre environnement à l'échelle cellulaire, j'ai répété ces études d'association en analysant un autre phénotype moléculaire : la réponse transcriptionnelle à l'activation du système immunitaire par différents composés ou pathogènes (reQTM, pour responseQTM). Enfin, mon second projet m'a permis d'analyser le niveau de spécificité cellulaire des associations détectées en utilisant des données de compte cellulaire par cytométrie de flux. Le recoupement entre eQTM et sites CpG associés à la génétique m'a permis d'identifier des situations où l'expression des gènes est potentiellement sous contrôle de facteurs génétiques et épigénétiques.

Finalement, le dernier axe de travail a été établi afin de clarifier ces situations où il est impossible d'identifier les relations causales entre génétique, épigénétique et expression des gènes. L'objectif de cette dernière facette de ma thèse était donc de développer et de

tester différentes méthodes pour établir la causalité de telles situations en utilisant des simulations, puis d'appliquer la méthode choisie aux trios gène-CpG-SNP précédemment établis. L'identification des associations gène-CpG a permis de caractériser plus précisément les régions du génome pour lesquelles la méthylation de l'ADN jouait un rôle actif sur l'expression des gènes, et donc d'améliorer notre compréhension des mécanismes sous-jacents.

Chapitre 5

Résultat 1 : Exploration des origines génétiques des différences populationnelles de méthylation de l'ADN et de leur impact causal sur la régulation des gènes de l'immunité

5.1 Contexte

Comme nous l'avons abordé dans les chapitres précédents, les variations épigénétiques peuvent avoir un impact sur la variabilité phénotypique. En particulier, la méthylation de l'ADN possède un rôle régulateur de l'expression des gènes via différents mécanismes que nous avons précédemment détaillés. De ce fait, la méthylation participe, via l'expression des séquences d'ADN, à notre interaction avec notre environnement. Par ailleurs nous avons montré comment de nombreux facteurs génétiques et environnementaux peuvent être responsables de modifications des profils de méthylation de l'ADN. Il en résulte que la méthylation de l'ADN se trouve à l'interface entre l'environnement et le génome, certaines études ont d'ailleurs directement étudié l'importance relative des mutations génétiques et des facteurs environnementaux sur l'établissement des profils de méthylation (Fagny et al., 2015).

De plus, un certain nombre d'études ont identifié des différences dans les profils de méthylation entre populations humaines (Fraser et al., 2012 ; Heyn et al., 2013 ; Moen et al., 2013 ; Carja et al., 2017). Et si certaines de ces études ont tenté d'identifier les causes ou les conséquences de telles différences, aucune ne s'est toutefois intéressée à ces deux aspects à la fois. Il a pourtant été montré, et nous l'avons détaillé dans les chapitres précédents, qu'il existe des différences inter-populationnelles dans les niveaux d'expression des gènes, en particulier des gènes de l'immunité, et que ces différences ont bien souvent des bases génétiques (Quach et al., 2016 ; Nedelec et al., 2016). On voit apparaître ici l'intérêt d'intégrer la nouvelle couche d'information que peut apporter la méthylation de l'ADN dans l'étude des différences populationnelles dans notre réponse à une activation de notre système immunitaire.

Dans cette étude nous avons tiré parti de données de séquençage d'ARNm, de profils de méthylation et de profils génétiques de monocytes primaires issus de 200 individus d'origines européenne ou africaine. Ces deux populations, chez qui des différences d'ex-

pression et de réponse transcriptionnelle à l'activation du système immunitaire avaient été détectées (Quach et al., 2016), nous ont servi de modèle pour tenter de démontrer l'apport d'inclure les marques épigénétiques dans l'étude de la régulation de la transcription. En effet, bien qu'une partie de ces différences avaient pu être attribuée à des variations inter-populationnelles de la séquence d'ADN, une part non négligeable de la variance de l'activité transcriptionnelle de ces cellules ne pouvait être attribuée à la seule variabilité génétique.

En premier lieu, nous avons identifié les différences de méthylation entre individus d'origines européenne et africaine, ainsi que leurs bases génétiques. Nous nous sommes ensuite intéressés à l'impact causal de la méthylation de l'ADN sur l'expression des monocytes dans un état non-stimulé, en combinant la variabilité génétique, la variabilité épigénétique et l'expression des gènes. Finalement nous avons répété ces dernières analyses en nous intéressant à la réponse transcriptionnelle des monocytes suite à leur activation par des composés mimant une infection bactérienne ou virale, ou par une souche du virus de la grippe.

5.2 Article 1

Husquin et al. Genome Biology (2018) 19:222
https://doi.org/10.1186/s13059-018-1601-3

Genome Biology

RESEARCH Open Access

Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation

Lucas T. Husquin[1,2,3], Maxime Rotival[1,2,3], Maud Fagny[4], Hélène Quach[1,2,3], Nora Zidane[1,2,3], Lisa M. McEwen[5], Julia L. MacIsaac[5], Michael S. Kobor[5], Hugues Aschard[3], Etienne Patin[1,2,3] and Lluis Quintana-Murci[1,2,3*]

Abstract

Background: DNA methylation is influenced by both environmental and genetic factors and is increasingly thought to affect variation in complex traits and diseases. Yet, the extent of ancestry-related differences in DNA methylation, their genetic determinants, and their respective causal impact on immune gene regulation remain elusive.

Results: We report extensive population differences in DNA methylation between 156 individuals of African and European descent, detected in primary monocytes that are used as a model of a major innate immunity cell type. Most of these differences (~ 70%) are driven by DNA sequence variants nearby CpG sites, which account for ~ 60% of the variance in DNA methylation. We also identify several master regulators of DNA methylation variation in *trans*, including a regulatory hub nearby the transcription factor-encoding *CTCF* gene, which contributes markedly to ancestry-related differences in DNA methylation. Furthermore, we establish that variation in DNA methylation is associated with varying gene expression levels following mostly, but not exclusively, a canonical model of negative associations, particularly in enhancer regions. Specifically, we find that DNA methylation highly correlates with transcriptional activity of 811 and 230 genes, at the basal state and upon immune stimulation, respectively. Finally, using a Bayesian approach, we estimate causal mediation effects of DNA methylation on gene expression in ~ 20% of the studied cases, indicating that DNA methylation can play an active role in immune gene regulation.

Conclusion: Using a system-level approach, our study reveals substantial ancestry-related differences in DNA methylation and provides evidence for their causal impact on immune gene regulation.

Keywords: Epigenetics, DNA methylation, Ancestry, Gene expression, Mediation, Immunity

Background

Individuals and populations display variable susceptibility to infectious diseases, chronic inflammatory disorders, and autoimmunity [1, 2]. Over the last decade, it has become clear that such disparities partly result from differences in the host genetic make-up, with an increasing number of genes being associated with varying abilities to fight infections at the individual and population level [3, 4]. Furthermore, population genetic studies have revealed that pathogen-driven selection has substantially impacted human genetic diversity [5, 6]. Because the mortality, and thus the selective pressure, imposed by pathogens have been paramount [7], human populations had to adapt to the different pathogenic environments they encountered around the globe, and genes involved in host defense are among the functions most strongly selected for by natural selection [5, 8–11]. While substantial evidence supports this hypothesis at the genetic level, we still know little about the degree of naturally

* Correspondence: quintana@pasteur.fr
[1]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France
[2]Centre National de la Recherche Scientifique (CNRS) UMR2000, 75015 Paris, France
Full list of author information is available at the end of the article

occurring epigenetic variation at the population level and how this may impact immune phenotypes.

As the immune system is the primary interface with the human pathogenic environment, the study of DNA methylation [12, 13] offers a unique opportunity to explore the interplay between the genome and environmental cues. DNA methylation can be affected by a range of external factors, such as nutrition, toxic pollutants, social environment, and infectious agents [14–19]. Furthermore, numerous studies have mapped DNA sequence variants associated with DNA methylation variation [20–28], i.e., methylation quantitative trait loci (meQTLs), and ~ 20% of the inter-individual variation in DNA methylation has been attributed to genetics [29, 30]. DNA methylation variation has also been associated with complex traits, including aging [31], body mass index [32], various cancers [33, 34], obesity [35], and autoimmune and inflammatory disorders [36, 37]. Yet, most studies of human epigenome variation, both in health and disease conditions, have focused on populations of homogeneous genetic ancestry, primarily of European descent.

A few studies, however, have reported that population differences in ancestry, habitat, or lifestyle affect DNA methylation, providing an initial assessment of the contribution of genetic factors and gene-environment (G × E) interactions to population-level epigenetic variation [38–44]. Yet, these studies investigated DNA methylation variation from virus-transformed lymphoblastoid cell lines or whole blood, so the differences observed could reflect, at least partially, epigenetic changes induced by cell immortalization or heterogeneity in blood cell composition that was not fully accounted for [45–47]. Thus, the extent of DNA methylation variation related to ancestry, and its genetic determinants, in a cellular setting relevant to immunity are far from clear.

A growing body of research has reported ancestry-related variation in terms of immune gene expression levels. Two recent studies found marked differences between individuals of African and European ancestry in their transcriptional responses to infectious challenges [48, 49] and showed that regulatory variants (i.e., expression quantitative trait loci, eQTLs) explain a substantial proportion of these population differences. Still, a large fraction of the variance in gene expression, both across individuals and populations, cannot be attributed to genetic factors and remains unexplained [48–55]. In this context, DNA methylation represents an additional, possible layer for variation in gene regulation [56]. The observed correlations between DNA methylation and gene expression levels can be positive and negative; in the canonical model, high levels of methylation at promoter regions are often associated with low gene expression, but elevated gene body methylation is also associated with active expression [28, 47, 57–60]. There is also increasing evidence that DNA methylation can play both passive and active roles in the regulatory interactions influencing gene expression, but the causality relationships between DNA methylation, gene expression, and genetic factors are not fully understood [19, 23, 56]. Furthermore, genetic variants associated with complex traits or diseases by genome-wide association studies (GWAS) often overlap both eQTLs and meQTLs, suggesting that disease risk can be mediated, directly or indirectly, by variation in DNA methylation [61–67].

Here, we aimed to broaden our understanding of the mechanistic links between ancestry-related differences in DNA methylation, genetic factors, and immune gene regulation. To do so, we build upon the EvoImmunoPop collection of primary monocytes originating from healthy individuals of African and European ancestry [48]. We profiled the DNA methylome of 156 donors, including 78 of each ancestry, using the high-resolution Infinium MethylationEPIC array, which captures methylation variation at more than 850,000 sites. This new dataset was combined with both genome-wide genotyping and whole-exome sequencing data, as well as with RNA-sequencing profiles from resting and stimulated monocytes with various immune stimuli, obtained from the same individuals. Such a system-level approach, integrating epigenetic, genetic, and transcriptional data, allowed us to assess the extent to which population-level variation in DNA methylation and its genetic determinants impact transcriptional activity related to immune responses.

Results

Population differences in DNA methylation profiles of primary monocytes

To assess population differences in DNA methylation of a purified innate immune cell type, we characterized DNA methylation variation at > 850,000 CpG sites across the genome, in monocytes originating from 156 male healthy volunteers: 78 of African descent (AFB, median age = 30.9 years) and 78 of European descent (EUB, median age = 25.9 years), all living in Belgium. Note that AFB individuals moved to Belgium between the ages of 6–45 years old (median age = 29 years). After normalization and filtering (see "Materials and methods"), we retained a final dataset of 552,141 methylation sites in the 156 individuals (Additional file 1: Figure S1). Principal component analysis (PCA) of DNA methylation clearly separated AFB and EUB along the first two PCs, which explained together 11.6% of the total variance (Fig. 1a). At a false discovery rate (FDR) = 1%, we identified 77,857 sites (14.1% of the total number) that presented a significant difference between AFB and EUB in their mean level of DNA methylation, after adjusting for age and surrogate variables. When restricting our analyses to CpGs that presented a mean difference > 5% (measured by the β value [68], see "Materials and methods"), we identified a total of 12,050

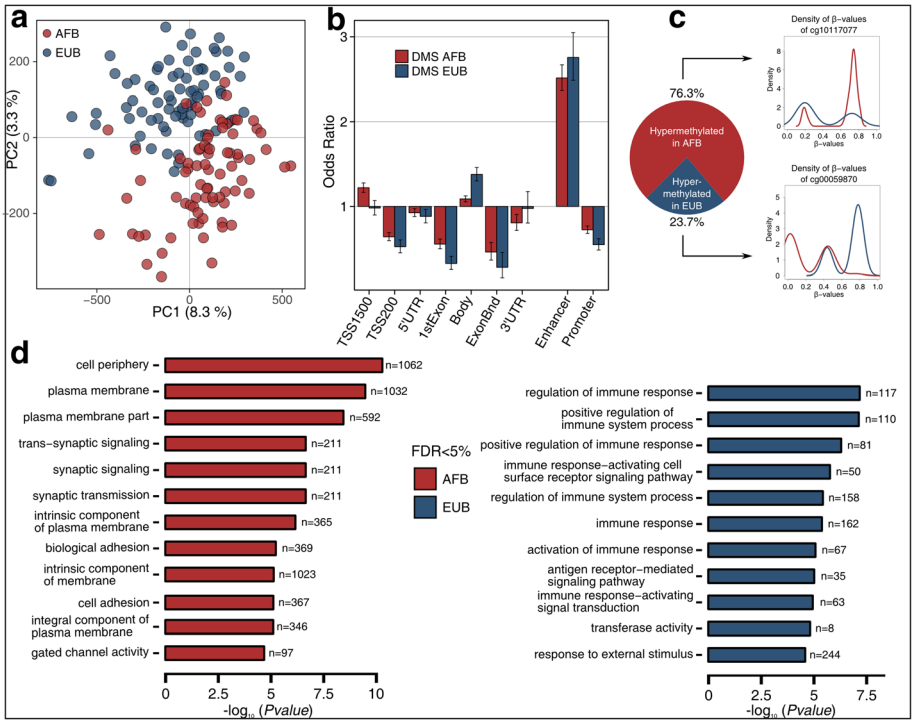

Fig. 1 Population differences in DNA methylation profiles. **a** Principal component analysis (PCA) of DNA methylation profiles for all 156 individuals. Red and blue circles represent African (AFB) and European (EUB) individuals, respectively. The proportions of variance explained by PC1 and PC2 are indicated. **b** Genomic location of differentially methylated sites (DMS), for CpG sites hyper-methylated in AFB (red) and in EUB (blue). Odds ratio and 95% confidence intervals are displayed for AFB-DMS and EUB-DMS, comparing their localization in different genomic locations as provided by Illumina (TSS1500, TSS200, 5′UTR, 1stExon, Body, Exon boundaries [ExonBnd], and 3′UTR), and in enhancer and promoter regions specifically detected in monocytes by ChromHMM phase 15 (see refs. [110, 111]). Odds ratios were computed against the general distribution of the 552,141 CpGs of our dataset. **c** Proportion of DMS that are hyper-methylated either in AFB (red) or in EUB (blue) individuals. The density of β values of one CpG site by category is given as an illustration of the population differences, with red and blue lines representing the methylation density in AFB and EUB, respectively. **d** Gene Ontology (GO) enrichment analyses of AFB- and EUB-DMS. For both groups, the top-GO categories reaching 5% FDR are shown, together with the number of genes per category and the $\log_{10}$-transformed FDR-adjusted enrichment *P* values

differentially methylated sites between populations (DMS) that mapped to 4818 genes. Because the age distributions of AFB and EUB individuals significantly differ (Wilcoxon *P* value = 10^{-4}; Additional file 1: Figure S2), and age might have a non-linear effect on DNA methylation [69], we also investigated with ANOVA the extent to which DNA methylation is non-linearly affected by age in our dataset. Our analyses showed that such effects had little to no impact on the population differences in DNA methylation detected (Additional file 2: Supplementary Note 1).

The genomic distribution of DMS, which were highly enriched in enhancer regions (odds ratio (OR) ~ 2.6, $P = 1.42 \times 10^{-224}$), was independent of the population where hyper-methylation was observed (Fig. 1b). However, of the 12,050 DMS, 76.3% were more methylated in AFB than in EUB, with respect to the observed 54% when considering all CpGs (Fisher's exact $P < 2.2 \times 10^{-16}$) (Fig. 1c). The corresponding genes were enriched in Gene Ontology (GO) categories related to cellular periphery and plasma membrane (Fig. 1d). The remaining 23.7%, which were hyper-methylated in EUB, were enriched in sites located in genes largely associated with immune response regulation and responses to external stimulus (Fig. 1c, d; Additional file 3: Table S1). These results cannot be explained by population differences in monocyte subpopulations (i.e., $CD14_{high}/CD16_{neg}$ [Classical], $CD14_{high}/CD16_{low}$ [Intermediate], and $CD14_{low}/CD16_{high}$ [Non-Classical]), as adding these subpopulations as covariates in the model did not alter our results (Additional file 1: Figure S3). Furthermore, we detected no CpG sites whose levels of methylation correlate significantly with monocyte subtypes (FDR = 5%), indicating that the effects of monocyte subpopulations on DNA methylation are negligible at the epigenome-wide level. Together, these analyses reveal genes and functions that present extensive differences in DNA methylation between individuals of African and European ancestry, in the context of primary monocytes.

Genetic factors drive most ancestry-related DNA methylation variation

We next examined the genetic determinants of the observed population differences in DNA methylation, and mapped methylation quantitative trait loci (meQTLs). We first tested for local associations between DNA methylation variation at CpGs and SNPs located within a 100-kb window (*cis*-meQTLs), using MatrixEQTL [70] (see "Materials and methods"). We set a 5% FDR threshold, considering one association per CpG site and using 100 permutations ($P < 1 \times 10^{-5}$). We adjusted for age, two surrogate variables (accounting for batch effects and unknown confounders, see "Materials and methods"), and the first two PCs of the genetic data (Additional file 1: Figure S4), to account for population stratification. To detect subtle effects, we merged all individuals and included ancestry as a covariate, but simultaneously, we analyzed the two populations separately to detect putative population-specific effects. For all subsequent analyses, we present the significant results of these two approaches combined, unless otherwise indicated.

We identified 69,702 CpGs associated with at least one genetic variant in at least one population (~ 12.6% of all sites, referred to as meQTL-CpGs). Given that multiple linked SNPs can be associated to the same CpG, we kept the best-associated SNP for each meQTL-CpG. However, we also used a fine mapping approach [51] to detect independent SNPs associated to each CpG (see "Materials and methods"). In doing so, we detected 9826 additional meQTLs (Additional file 1: Figure S5), providing a more thorough view of the contribution of proximate genetic variants to DNA methylation variation. The median distance between a CpG and its associated SNP was ~ 3.8 kb (Additional file 1: Figure S6), supporting the close genetic control of DNA methylation [22, 28, 41, 65]. Furthermore, we found a 2.2-fold enrichment of meQTL-CpGs in enhancers ($P < 1 \times 10^{-326}$), a trend that was even more pronounced for meQTLs associated with population differences in DNA methylation (meQTL-DMS; OR ~ 2.8, $P = 6.8 \times 10^{-317}$, Additional file 1: Figure S7).

Focusing on ancestry-related differences, we observed that ~ 70.2% of DMS harbor a significant meQTL, with respect to the 12.6% detected genome-wide (Fisher's exact $P < 2.2 \times 10^{-16}$; Fig. 2a). These meQTLs were found to account, on average, for ~ 58% of the observed population differences in DNA methylation (Additional file 1: Figure S8, see "Materials and methods"). Furthermore, meQTLs presented opposite effects on DNA methylation as a function of population differences in allelic frequency, i.e., a derived allele at higher frequency in Africans was generally associated with high levels of DNA methylation, while a derived allele at higher frequency in Europeans was primarily associated with low DNA methylation (Fig. 2b). This observation provides a genetic explanation for the unbalanced patterns of hyper-methylation, observed at DMS, between Africans and Europeans (Fig. 1c).

Local meQTLs can, a priori, lead to population differences in DNA methylation following two main models: (i) the meQTL has a similar effect in both populations but present different allelic frequencies (Fig. 2c), or (ii) the meQTL is present at similar frequencies but display population-specific effects, revealing more complex interactions (Fig. 2d). We therefore investigated the population specificity of the 69,702 meQTL-CpGs detected using a model selection approach (see "Materials and methods"). We found 2868 (4.1%) significant population-specific effects (1337 AFB-specific and 1531 EUB-specific), suggesting the occurrence of G × E or G × G effects.

Ancestry-related meQTLs are enriched in associations with complex traits and diseases

Given that a large fraction of genetic variants identified by GWAS are thought to act by affecting gene regulation [71–74], we investigated the putative functional impact of the detected meQTLs on ultimate complex phenotypes. In practice, we searched for enrichments in GWAS hits among our set of 79,528 meQTLs, correcting for linkage disequilibrium (see "Materials and methods"). Focusing on the 17 parental classes of the Experimental Factor Ontology (EFO) classification [75], we found that meQTLs were enriched in significant hits for all these functional categories (Additional file 1: Figure S9, OR ~ 2.1–5.5, $P < 4.1 \times 10^{-10}$). Stronger enrichments were detected for meQTLs associated with population differences in DNA methylation (OR ~ 2.7–9.8, $P < 2.9 \times 10^{-3}$), in particular for phenotypes related to hematological measurements, neurological disorders, immune system disorders, inflammatory measurements, and digestive system disorders (Fig. 2e).

Because DNA methylation and meQTLs have been shown to be largely cell or tissue dependent [23, 76–81], we next searched for the specific traits that account for the signals detected at the parental category "immune system disorder", given our focus on primary monocytes. We found that meQTLs overlapped variants associated with diseases such as osteoarthritis, psoriasis, systemic lupus erythematosus, inflammatory skin disease, or type 1 diabetes (Additional file 1: Figure S10). For example, the meQTL SNP rs629953 presents markedly different frequencies between AFB and EUB (DAF AFB 7.5% versus DAF EUB 62%), leading to variable population-level DNA methylation at *TNFAIP3* (cg06987098), and has been associated with psoriasis susceptibility [82, 83]. Together, our analyses support that complex traits and variable DNA methylation are pleiotropically associated with genetic variation [39, 60, 63, 64], but extend these

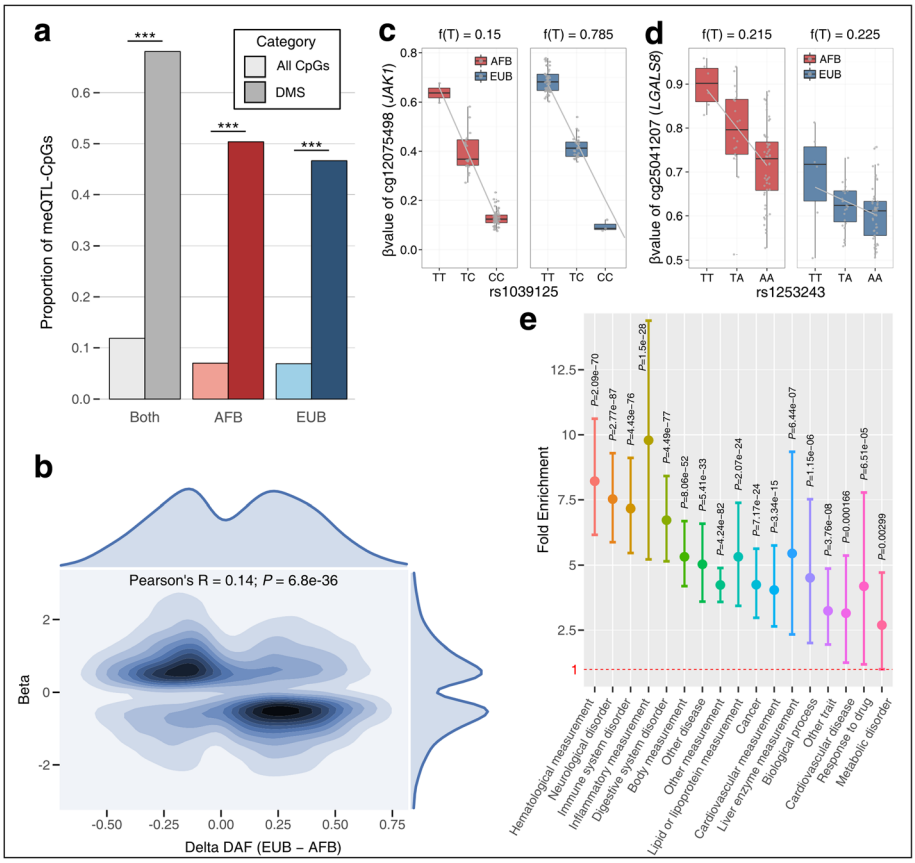

Fig. 2 Genetic control of population differences in DNA methylation levels. **a** Proportions of CpGs and DMS associated to genetic variants identified in the three meQTL studies: merging the two populations (gray shades), mapping in AFB only (red shades) and in EUB only (blue shades). For each mapping, proportions among all 552,141 tested CpG sites and among DMS are indicated in light and dark colors, respectively. ***Fisher's exact $P < 2.2 \times 10^{-16}$. **b** Contour plot of meQTL effects on DMS as a function of their difference in derived allelic frequencies (DAF) between populations. For each of the 8459 DMS for which we detected at least one meQTL, we used a kernel density estimation to draw the contour plot of the effect of the derived allele of the meQTL onto methylation (beta, *Y* axis) according to the ΔDAF ($DAF_{EUB} - DAF_{AFB}$, *X* axis). The coefficient and *P* value of Pearson's correlation test are displayed. The marginal distribution of the two variables is displayed: top for ΔDAF, and right for beta. **c**, **d** Examples of meQTLs detected in this study. Boxplots represent the distribution of *β* values as a function of genotype, for AFB (red) and EUB (blue) individuals. The minor allele frequency of each meQTL is presented for each population on the top. Gray lines indicate the fitted linear regression model for *β* value~genotype for each population. **e** Fold enrichment of meQTLs associated with DMS in GWAS hits. For each of the 17 parental EFO categories, the fold enrichment, the 95% confidence intervals obtained by bootstrap, and the associated *P* values are shown

associations to variants affecting ancestry-related epigenetic variation in the context of an innate immunity cell type.

Exploring the distant genetic control of DNA methylation variation

We subsequently searched for the effects of distant genetic variants on DNA methylation variation (*trans*-meQTLs). To limit the burden of multiple testing, and because *trans*-meQTLs are enriched in *cis*-eQTLs for genes encoding transcription factors (TF) [65], we focused on two non-independent subsets of genetic variants: (i) the 4037 SNPs detected as *cis*-eQTLs for one of 600 TF-coding genes and, more generally, (ii) the 73,561 SNPs located in the vicinity (± 10 kb) of the TSS of these genes. Only associations for which the SNP-CpG distance was higher than 1 Mb were considered, at an FDR of 5% ($P < 1 \times 10^{-9}$). Given the generally low power to map *trans*-associations, we performed this analysis by considering all individuals together and including ancestry as a covariate.

We identified 133 CpG sites associated with at least one distant SNP, for a total of 672 *trans*-meQTLs that involved 91 independent loci (Additional file 4: Table S2). Among these, we detected a number of hubs of distant genetic control of DNA methylation variation, including six TFs (*ZNF429*, *CTCF*, *FOXI1*, *ZBTB25*, *MKL2*, and *NFATC1*) where local genetic variation was associated with at least 10 different CpGs in *trans*. Highlighting one pertinent example, a single genetic

variant (rs7203742) nearby *CTCF*—encoding a transcriptional regulator with 11 highly conserved zinc-finger domains—controls the degree of DNA methylation at 30 CpG sites, ~ 29.4% of all CpGs regulated in *trans*. Furthermore, of the 21 *trans*-regulated CpGs that were detected as DMS, 12 were controlled by the same *CTCF* variant. That this variant (T → C) presents high levels of population differentiation (DAF AFB 24% vs. EUB 88%, $F_{ST} = 0.59$ in the 1% of the genome-wide distribution) suggests the action of positive selection targeting the derived allele in Europeans. This observation makes of *CTCF* not only a master regulator of DNA methylation, as previously observed [65], but also an important contributor to differences in DNA methylation between human populations.

Dissecting the mechanistic relationships between DNA methylation and gene expression

We leveraged the availability of RNA-sequencing data from the same individuals [48] to obtain new insights into the mechanistic relationships between DNA methylation and gene expression variation, in African and European individuals. We associated the levels of expression of 12,578 genes in primary monocytes with those of DNA methylation at CpGs located within 100 kb of their TSS, for a total of 513,536 CpG sites. Associations were considered significant if they passed a *P* value threshold determined using 100 permutations (FDR = 5%, $P < 5 \times 10^{-5}$) (see "Materials and methods").

We identified 1666 CpGs whose levels of DNA methylation were associated with gene expression (eQTMs), for a total of 811 genes (eQTM-genes) associated with at least one CpG in one population group (Additional file 5: Table S3). The KEGG pathways associated with eQTM-genes contained a large number of immune-related pathways, providing a link between DNA methylation and gene expression in the context of immunity (Fig. 3a). When investigating the population specificity of the 811 eQTMs (see "Materials and methods"), we detected 93 significant population-specific effects (43 AFB-specific and 50 EUB-specific). The majority of these cases (80 out of 93) corresponded to genes whose eQTMs were also under genetic control, suggesting, again, the occurrence of G × G or G × E interactions.

Based on current genomic annotations, eQTMs were mostly negatively correlated to gene expression (69.5% vs. 30.5%, see also refs. [23, 28, 65, 84, 85]). Negatively correlated sites were strongly enriched in enhancers (OR~ 2.6, $P = 6.6 \times 10^{-59}$) (Fig. 3b), highlighting their major role in transcriptional regulation [86–88]. In addition, we found a slight excess of negative associations in promoters (OR ~ 1.2, $P = 1.8 \times 10^{-2}$) and nearby TSS (TSS1500) (OR ~ 1.4, $P = 7.2 \times 10^{-13}$), as expected following the canonical model. Conversely, positive associations were enriched in sites located nearby UTRs, particularly 3′-UTR (OR ~ 1.8, $P = 8.4 \times 10^{-5}$) [89], but depleted in sites located in promoters (OR ~ 0.6, $P = 1.1 \times 10^{-4}$) (Fig. 3b). Furthermore, we found that eQTMs were strongly enriched in DMS (OR ~ 11.8, $P < 1.93 \times 10^{-216}$) and, importantly, in meQTL-CpGs (OR ~ 33.2, $P < 1 \times 10^{-326}$) (Fig. 3c). Together, these observations indicate that DNA methylation variation, in particular at sites that are differentially methylated across populations (DMS), is much more likely to be under genetic control when associated with gene expression differences (eQTMs), than random CpG sites.

Exploring the underlying causality between regulatory loci and gene expression

Because the respective roles of genetic and epigenetic factors in transcriptional regulation are not fully understood [56], we next mapped eQTLs (FDR = 5%, see "Materials and methods") to identify the cases where DNA methylation, gene expression, and genetic variants show significant associations between all pairs (Additional file 1: Figure S11). We thus obtained 552 trios, each of them consisting of one gene, one to various CpGs and one to various SNPs (containing 68.1% of the genes detected in the eQTM mapping). This suggested potential, causal relationships between these variables—a latent, though challenging, question in epigenetics. To infer causality between regulatory loci (i.e., eQTMs and eQTLs) and gene expression variation for these specific trios, we first used an elastic net model to build two intermediate variables measuring (i) DNA methylation variability attributable to genetics for the trios presenting more than one SNP and (ii) gene expression variability attributable to DNA methylation for the trios presenting more than one CpG (see "Materials and methods").

We used a Bayesian approach [90] to assess potential causal effects of a mediating variable *M* (DNA methylation) on the relationship between an independent variable *X* (genetics) and a dependent variable *Y* (gene expression) [91]. When comparing the performance of this method with that of an approach based on partial correlations, using simulated data and various genomic scenarios, we found similar results between the two approaches in terms of sensitivity and specificity (Fig. 4a, b; Additional file 1: Figure S12; see "Materials and methods"). We then ran the mediation analysis on each trio, adjusting for regular covariates (age and surrogate variables), but also for the fourth and second PCs of gene expression and DNA methylation, respectively. The latter covariates were added because they likely capture potential confounding factors inducing correlation between DNA methylation and expression, which would violate the assumption of the causal inference model (Additional file 1: Figure S13). Note that reverse

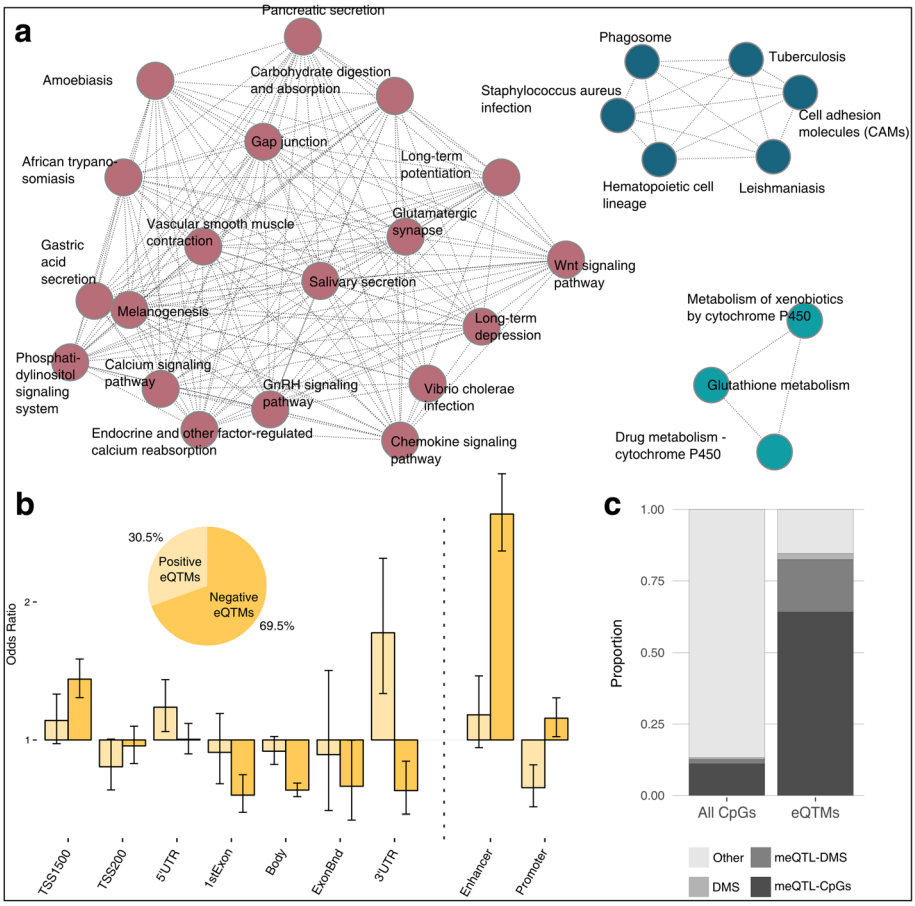

Fig. 3 Correlations of DNA methylation with gene expression. **a** Networks of KEGG pathways of genes detected in the eQTM mapping. **b** Genomic location of eQTMs, for positively and negatively associated CpG sites (light and dark yellow, respectively). Odds ratio were computed against the general distribution of the 552,141 CpGs from our dataset. The distribution of eQTMs according to the direction of their effect on gene expression is shown. **c** Proportions of different groups of CpG sites in all tested sites (left panel) and among the detected eQTMs (right panel)

causation was found to be unlikely in our experimental setting and was thus not considered in our analyses (Additional file 2: Supplementary Note 2).

At FDR = 5%, we identified 165 genes where the genetic control of expression levels was mediated by DNA methylation (i.e., $\alpha \times \beta$ was significantly different from zero, Fig. 4a), in at least one population. Remarkably, in 66 of these cases, mediation occurred through CpG sites that are differentially methylated across populations (DMS) (Additional file 6: Table S4). The proportion of mediated genes whose expression was positively and negatively correlated to DNA methylation was similar, ranging from 26 to 31% (Fig. 4c). Expectedly, we found that, among mediated genes, DNA methylation explained a significantly higher proportion of the variance of gene expression than genetics (mean $R^2 = 23.4\%$ versus 15.4%, respectively; Wilcoxon $P = 3.3 \times 10^{-11}$), in contrast with the 387 non-mediated cases where we observed the opposite trend (Wilcoxon $P = 7.8 \times 10^{-37}$) (Fig. 4d).

We also found that CpG sites mediating gene expression were preferentially located in enhancers (OR ~ 2.5, $P = 4.0 \times 10^{-21}$), highlighting again the major role of these regions in epigenetic regulatory mechanisms [92–94]. These CpGs were depleted in promoters (OR ~ 0.7, $P = 1.4 \times 10^{-2}$), which were otherwise enriched in non-mediating CpGs (OR ~ 1.3, $P = 5.9 \times 10^{-3}$). Notably, 86.6% of mediating CpGs fell directly into a TF-binding site (TFBS), with respect to the expected 76.9% at the genome-wide level (OR ~ 1.9, Fisher's exact $P = 8.64 \times 10^{-7}$). This result suggests that DNA methylation might actively regulate transcriptional activity through the modulation of TF binding, a hypothesis that requires experimental validation.

Interestingly, among mediated cases, we found key genes of the immune response, such as *NLRP2*, *RAI14*, *NCF4*, or *ICAM4*, and genes with functions related to transcriptional activity, encoding zinc-finger proteins (Additional file 6: Table S4). This suggests a more extensive role of DNA methylation in regulating gene

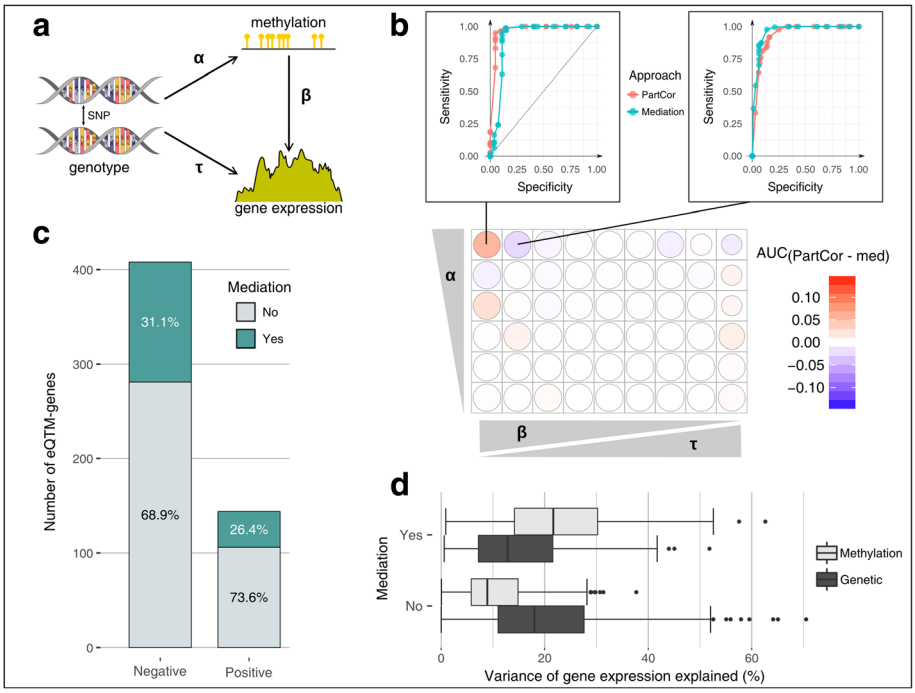

Fig. 4 Inference of the causal effects of DNA methylation on gene regulation. **a** Representation of a simulated scenario, with the three varying parameters (*α*, *β*, and *τ*). **b** Comparison of the mediation analysis (med) with a partial correlation approach (PartCor) using a range of different simulated parameters for *α* (0.3–0.8), *β* (0.9–0.1), and *τ* (0.1–0.9). Note that the parameter range simulated for *β* and *τ* was adjusted so that we kept 75% of the variance unexplained (random noise parameter $\gamma = 0.25$). The difference of the area under the curve (AUC) between the two approaches is represented with different shades of red and blue. The sizes of the circles are proportional to the mean AUC of the two approaches. Two examples of the ROC curves are shown in the upper part of the figure. **c** Number of mediated and non-mediated eQTM-genes for negative and positive associations between DNA methylation and gene expression. The percentages of these two categories are also indicated. **d** Proportion of variance of gene expression explained by DNA methylation (light gray) and genetics (dark gray), in mediated and non-mediated cases

expression than the local associations described here, through the regulation of DNA-binding protein activity.

Impact of immune perturbation on genetic and epigenetic interactions

Finally, we sought to understand how DNA methylation variation at the basal state affects transcriptional responses to immune activation. We used RNA-sequencing data, obtained from the same individuals, after exposure to various stimuli: LPS activating TLR4 and Pam3CSK4 activating TLR1/2, both pathways sensing bacterial components, R848 activating TLR7/8, predominantly sensing viral nucleic acids, and influenza A virus (IAV) [48]. We then mapped response-QTMs (reQTMs) using fold changes in gene expression between non-stimulated and stimulated states, for all genes expressed in either condition (see "Materials and methods").

We found 230 unique genes whose response to immune activation was associated with DNA methylation in at least one condition; most associations were context-specific, with only 7 genes detected in all conditions (Fig. 5a; Additional file 5: Table S3). Furthermore, a 2.5-fold increase was observed in the number of reQTM-genes detected upon activation with viral-stimuli (R848 and IAV; 197 unique genes) with respect to those detected for bacterial ligands (LPS and Pam3CSK4; 78 unique genes) (Fig. 5a). For example, we detected a reQTM upon R848 stimulation for *CARD9* in EUB and *CD1D* upon IAV infection in AFB, both genes known to play an important role in host defense (Fig. 5b, c). Despite reQTMs and eQTMs present a similar genomic distribution (Additional file 1: Figure S14), we observed an important shift towards positive associations between DNA methylation and transcriptional responses, in particular to TLR ligands (Fig. 5d). This shift was mainly accounted for by reQTMs that present the strongest associations between DNA methylation and gene expression in the non-stimulated condition (Additional file 1: Figure S15), corresponding to 109 genes (47% of the total). This contrasts with the canonical model of negative associations primarily observed at reQTMs presenting the strongest associations at the stimulated state, corresponding to 131 genes (57% of the total). Note that 10 genes were associated with reQTMs of both groups.

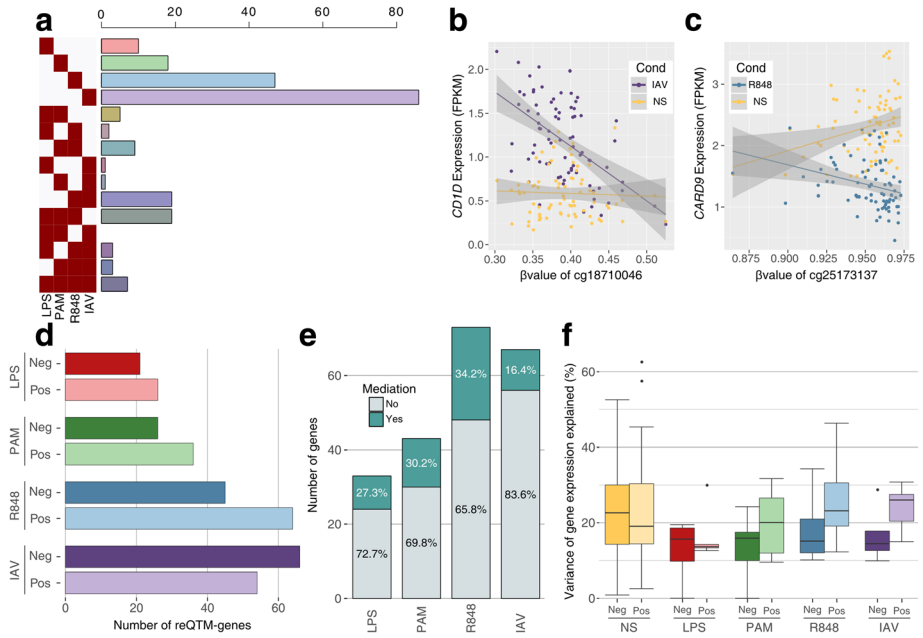

Fig. 5 Effects of DNA methylation on transcriptional responses to immune stimulation. **a** Number of genes harboring reQTMs in single conditions or combinations of stimulations. **b**, **c** Examples of reQTMs detected in this study. Lines indicate the fitted linear regression model, and gray shades the 95% confidence intervals of these models. **b** The distribution of the expression values of *CD1D* at the non-stimulated (yellow) and after IAV infection (purple) is plotted as a function of β values, for AFB individuals only. **c** The distribution of the expression values of *CARD9* at the non-stimulated (yellow) and upon R848 stimulation (blue) is plotted as a function of β values, for EUB individuals only. **d** Number of reQTM-genes by condition and according to the direction of their association with DNA methylation. **e** Number of mediated and non-mediated reQTM-genes per stimulation condition. The percentages of these two categories for each condition are also indicated. **f** Proportion of variance of gene expression explained by DNA methylation, among negative (dark colors) and positive (light colors) associations, in mediated cases

To explore causal mediation effects of DNA methylation in the context of immune activation, we mapped response-QTLs (see "Materials and methods"). Following our previous rationale (Additional file 1: Figure S11), we identified 141 trios (61.3% of the 230 reQTM-genes, Additional file 6: Table S4). At FDR = 5%, we detected 40 genes (28.4%) where the genetic control of their transcriptional response was mediated by DNA methylation (Fig. 5e). Although non-significant, we found a higher proportion of mediation for genes whose response was positively associated with DNA methylation, as compared to negative associations, in particular for viral challenges (OR ~ 2.0; Fisher's exact $P = 0.33$) (Additional file 1: Figure S16). Among mediated genes in the viral conditions, the proportion of gene expression variance explained by DNA methylation was higher for positive than for negative associations, again at odds with the non-stimulated condition (Fig. 5f). More generally, our analyses illustrate the value of mapping reQTMs and studying the underlying patterns of causality, to uncover mechanisms that might explain disparities in the way individuals and populations respond to immune activation.

Discussion

Our population epigenetic results, obtained in the setting of an innate immunity cell population, demonstrate extensive differences in DNA methylation profiles between two populations that differ in their genetic ancestry but share the same present-day environment. Such population differences were observed at the epigenome-wide level (explaining ~ 12% of the total variance in DNA methylation) and involved 12,050 sites that were mostly located in genes with functions related to cell periphery or immune response regulation. Previous studies have searched for ancestry-related differences in DNA methylation in various human populations and cell types [16, 38–41, 43, 95]. Although comparisons across studies are complicated by differences in experimental settings and statistical thresholds used to detect ancestry-associated CpG sites, these range from 299 between Caucasian- and Asian/mixed-descent individuals living in Canada [16] to 36,897 between European CEU and African YRI [39]. An interesting insight that can be drawn from our analyses is that genes involved in the activation and regulation of immune responses tend to present higher levels of DNA methylation in individuals of European ancestry, with respect to those of African

ancestry, mostly owing to genetic control. That up to 16% of immune-related genes that are hyper-methylated in Europeans are also differentially expressed between populations [48] could provide a mechanistic explanation for the ancestry-related differences in transcriptional responses to bacteria reported in macrophages, where European ancestry is associated with lower inflammatory responses [49].

Although variation in past environmental exposures and socioeconomic factors may contribute to population differences in DNA methylation, we found that 70% of differentially methylated sites between African and European ancestry groups were associated with at least one meQTL. This indicates that population differences in DNA methylation are mostly driven by DNA sequence variants [38, 40–42]. In some cases, a single genetic variant can account for important population differences at multiple CpG sites, as attested by the *trans*-meQTL we detected at *CTCF*, whose local genetic variation has been shown to alter distant DNA methylation patterns in whole blood [65]. We show that a *CTCF* variant (rs7203742) regulates DNA methylation of 30 distant CpGs, 40% of which are differentially methylated between populations. We also found that all *CTCF trans*-regulated CpGs fall within a TFBS, confirming our initial hypothesis about the mechanism by which a genetic variant might alter DNA methylation at a distant CpG site. Interestingly, 9 out of the 30 *CTCF trans*-regulated CpGs fall within a TFBS of *CTCF*, while the remaining 21 fall within a TFBS specific to other TFs such as *YY1*, *ESR1*, or *ZNF143*. This observation is consistent with a model of pioneer transcription factor activity [96] and suggests that *CTCF* acts as a pioneer factor that will generate changes in chromatin state that, in turn, will become accessible for binding of secondary factors.

At the genome-wide level, we find that the quantitative impact of DNA methylation on gene expression variation is lower than that reported by some previous studies, possibly reflecting differences in experimental settings and statistical power (e.g., cell types and sample sizes) [23, 65, 84, 89]. For example, a study of 204 healthy newborns detected substantial variation across tissues in the number of genes whose expression levels were associated with DNA methylation, ranging from 596 in fibroblasts to 3838 in T cells [23]. We detected, at the non-stimulated state, 811 eQTM-genes (6% of the total number of expressed genes), a figure that drops to 230 for reQTM-genes across stimulation conditions. However, a limitation of our study is that we measured DNA methylation at the basal state, while gene expression was obtained after 6 h. Studies including a more comprehensive range of epigenetic marks obtained at different time points—in different cell types and tissues originating from individuals of various ancestries—are needed to more precisely understand the interplay between these regulatory elements and quantify their respective roles in the regulation of transcriptional activity.

The detected eQTMs were found to be drastically enriched in genetic control (OR ~ 33.2, $P < 1 \times 10^{-326}$, Fig. 3c), which highlights the coordinated action of genetic and epigenetic factors in driving gene expression variation but raises questions about the causal role of DNA methylation [56]. Despite cautious interpretation of causality in mediation analyses is required [97], our analysis provides a first estimate of the potential direct role of DNA methylation in regulating transcriptional activity, in both resting and stimulated monocytes. At the non-stimulated state, we find that $\sim 20\%$ of eQTM-genes show evidence of a causal mediation effect of DNA methylation. Although a similar extent of mediation was found upon immune stimulation ($\sim 17\%$), we detected specific patterns upon treatment with viral challenges, where a higher occurrence of positive associations was observed among mediated cases. These findings mostly reflected cases where high levels of DNA methylation were associated with low gene expression in the non-stimulated condition, thus requiring stronger responses to reach high levels of gene expression upon cell perturbation. These trends suggest a major, direct, and context-specific role of DNA methylation in the regulation of immune responses, whose complexity requires further investigation.

Finally, we found that meQTLs, in particular those associated with ancestry-related differences, are enriched in GWAS hits related to immune disorders. This suggests that DNA methylation has an important impact on the cellular activity of monocytes and ultimately affect phenotypic outcomes. Nonetheless, a large fraction of the variance of DNA methylation and gene expression remains unexplained. Additional work is needed to quantify the relative impact of genetic, epigenetic, environmental, and lifestyle factors in driving variation of DNA methylation and gene expression, both in resting and stimulated cells. Furthermore, although the causal mediation analyses presented in this study reinforce the notion that DNA methylation can play a direct role in regulating gene expression in humans [23, 98], monitoring the kinetics of variation in DNA methylation and gene expression after exposure to different infectious agents will broaden our understanding of the interplay between these molecular phenotypes and their impact on endpoint phenotypes.

Conclusion

Our study reveals extensive variation in DNA methylation profiles between individuals and populations, with ancestry-related differences being mostly explained by genetic variation. It also suggests that DNA methylation can have a direct, causal impact on the transcriptional

activity of primary monocytes, providing new insight into the nature of the host factors that drive immune response variation in humans.

Materials and methods

Sample collection and monocyte purification

The EvoImmunoPop collection consists of 156 individuals (males between 20 and 50 years old, mean 31.5 years old) from two different ancestries (78 of European and 78 of African descent), who were recruited at the Center for Vaccinology from the Ghent University Hospital (Ghent, Belgium) [48]. For each participant, 300 ml of whole blood was collected into anticoagulant EDTA-blood collection tubes and peripheral blood mononuclear cells (PBMCs) were purified using Ficoll-paque density gradients (#17-1440-03, GE Healthcare). Monocytes were positively selected from purified PBMCs using magnetic CD14 microbeads (#130-050-201, MiltenyiBiotec), as per manufacturer's instructions. All samples had a monocyte purity higher than 90% with a mean value of 97%.

DNA methylation profiling and data normalization

Genomic DNA was extracted from the monocyte fraction using a phenol/chloroform protocol followed by ethanol precipitation. The DNA was then bisulfite converted, and BC-DNA was then processed using the Illumina Infinium MethylationEPIC BeadChip Kit (Illumina, San Diego, CA) to obtain the methylation profile of each individual at more than 850,000 CpG sites genome-wide.

In total, 184 samples were hybridized with the EPIC array, including 172 unique samples and 12 technical replicates. We removed any technically unreliable probes: (i) potentially cross-hybridizing probes (83,635 probes), (ii) those located on the X and Y chromosomes (17,229 probes), and (iii) probes overlapping SNPs that present a frequency higher than 1% in at least one of the studied populations (206,998 probes). These SNPs were chosen based on our own genotyping dataset, as well as on the 1000 Genomes project [99]. To control for the quality of the probes and samples, we filtered out individuals with > 5% of probes associated with a detection P value > 10^{-3}, and then, probes with a detection P value > 10^{-3} in one or more individuals (6833 probes). Following this filtering process, 552,141 of the original 866,836 sites on the array were retained.

We calculated methylation levels from raw data, using the R Bioconductor lumi package [100]. Given that the M value has been shown to provide better detection sensitivity than β values at extreme levels of modification [68], we used the M value to run all statistical analysis unless otherwise stated. Note that in some instances of the text and figures, β values are reported for ease of clarity and interpretation. M values were then adjusted for background noise with the normal-exponential using out-of-band probes (noob) from the R Bioconductor minfi package [101]. Next, normalization for color bias was performed using *lumiMethyC* with the "quantile" method, and for methylated/unmethylated intensity variation using the *lumiMethyN* with the "ssn" method [100]. Finally, we corrected for technical differences between type I and type II assay designs, by performing beta-mixture quantile normalization [102]. To correct for known batch effects and potential hidden confounders, we used the *sva* function from the sva Bioconductor package [103] with age as a variable of interest. Additionally, five EUB samples were removed because they presented an excess of hemimethylated sites, leaving 89 EUB and 78 AFB samples. To obtain equal power in the two studied populations, we down-sampled the European group to 78 samples by randomly removing 11 EUB samples, for an overall final cohort of 156 individuals.

Extraction of differentially methylated sites (DMS)

To detect CpG sites presenting statistically different levels of DNA methylation between AFB and EUB, we fitted a linear regression model for each CpG site: M value ~ population + age + two surrogate variables + error, and next applied an empirical Bayes smoothing to the standard errors using the R Bioconductor limma pipeline [104]. P values were adjusted using the Benjamini and Hochberg method. DMS were extracted using a threshold of adjusted P value (< 0.01) and a difference in the mean β value of each population $|\Delta\beta| > 5\%$.

Mapping of methylation quantitative trait loci (meQTLs)

All individuals were genotyped for a total of 4,301,332 SNPs on the Illumina HumanOmni5-Quad BeadChips and went through whole-exome sequencing with the Nextera Rapid Capture Expanded Exome kit, on the Illumina HiSeq 2000 platform, with 100-bp paired-end reads. Details of the processing of genotyping and whole-exome sequencing data, together with imputation using the 1000 Genomes Project imputation panel [99], are reported in ref. [48]. For the meQTL mapping, we filtered out SNPs with a minor allele frequency < 5% in the populations studied and kept a final dataset of 10,278,745 SNPs (i.e., corresponding to the merged genotyping and whole-exome sequencing dataset after imputation; 8,913,090 SNPs in Africans and 6,178,808 SNPs in Europeans). Age, PC1 and PC2 of the genotype matrix, and two surrogate variables, as identified with the *sva* R package, were used as covariates in the linear model.

We mapped meQTLs using the statistical framework implemented in the MatrixEQTL R package [70]. For local associations (i.e., distance SNP-CpG ≤ 100 kb), we performed two independent mappings using (i) the direct linear model from the MatrixEQTL pipeline and (ii) a Kruskal-Wallis rank test. Associations were considered

significant when passing the 5% FDR threshold in both mappings. Two models were considered: merging all individuals and including a binary variable adjusting for ancestry or keeping the two populations separately. To detect all possible independent SNPs regulating methylation at a single CpG site in *cis*, we regressed out genotypes of all primary *cis*-meQTLs and then performed *cis*-meQTL mapping on the regressed methylation data to find secondary *cis*-meQTLs. We repeated this process in a stepwise fashion until no additional independent *cis*-meQTLs were detected. This allowed us to refine our local meQTL mapping by detecting all possible independent SNP-CpG associations.

For distant, *trans*-acting associations (i.e., distance between SNP and CpG ≥ 1 Mb or on different chromosomes), we restricted our analysis to SNPs located in the vicinity of transcription factor (TF) coding genes, to limit the burden of multiple testing. Specifically, we selected (i) all SNPs located less than 10 kb to the TSS of any expressed TF in our dataset and (ii) SNPs detected as *cis*-eQTLs for these TFs. For each SNP, we only investigated CpG sites that mapped at least 1 Mb from the SNP or located on other chromosomes, using a Kruskal-Wallis rank test.

For both *cis*- and *trans*-meQTLs, FDR was computed by mapping meQTLs on 100 datasets with the *M* values permuted within each population. We then kept, after each permutation, the most significant *P* value per CpG site, across populations (probe-level FDR). Finally, we computed the FDR associated with different *P* value thresholds for *cis* or *trans*, and subsequently selected the *P* value threshold that provided a 5% FDR: $P = 1 \times 10^{-5}$ and $P = 1 \times 10^{-9}$ for *cis*- and *trans*-meQTLs, respectively.

Investigating the genetic basis of population differences in DNA methylation

We aimed at identifying the proportion of the population differences in DNA methylation that was accounted for by genetic variability. To do so, for the 8459 DMS that were associated with at least one meQTL, we computed the following ratio:

$$ExpDiff = \frac{\beta \times \Delta DAF}{\Delta Meth}$$

with β reflecting the effect of the derived allele of the meQTL on methylation, ΔDAF the difference in allelic frequencies between Europeans and Africans ($DAF_{EUB} - DAF_{AFB}$), and ΔMeth the observed difference in the mean levels of DNA methylation between European and African individuals ($\overline{Meth_{EUB}} - \overline{Meth_{AFB}}$).

Note that this ratio is not bound to [0:1], as the effect of genetics onto the overall population differences in DNA methylation can be counteracted by opposite effects of independent origins (e.g., environmental factors or non-detected independent genetic effects).

Detecting population-specific meQTLs

We aimed at refining our meQTL mapping by detecting population-specific meQTL effects (i.e., SNPs present at similar frequencies in both populations but having different effect sizes on DNA methylation between populations). To do so, we used a Bayesian model selection approach to identify specific and shared effects for each of the 69,702 CpGs that we detected as being associated with at least one genetic variant. Specifically, for each CpG-SNP pair, we computed the likelihood of three models:

$$lm(Meth \sim SNP + Pop) \quad \text{(i)}$$

$$lm(Meth \sim SNP_{EUB} + Pop) \quad \text{(ii)}$$

$$lm(Meth \sim SNP_{AFB} + Pop) \quad \text{(iii)}$$

with SNP_{EUB} coded 0,1,2 in EUB individuals and 0 in AFB individuals, and SNP_{AFB} coded 0,1,2 in AFB individuals and 0 in EUB individuals. We next calculated the posterior probability of each model assuming that all models are equally likely a priori. We then set a threshold of 0.9 to consider one of the models as supported by the data. Thus, a meQTL is classified as EUB-specific if the posterior probability of model (*ii*) is higher than 0.9, or AFB-specific if the probability of model (*iii*) is higher than 0.9.

GWAS enrichment analyses

We used the NHGRI GWAS catalog [105] to first select all significant SNPs that were significantly associated with a complex trait or disease at a $P < 1 \times 10^{-8}$. Using this set of GWAS hits, we next extracted all SNPs in LD with each of these hits ($R^2 > 0.8$) and classified the resulting final set of 166,248 SNPs according to their parental Experimental Factor Ontology (EFO) term [75].

We then selected all meQTLs in our dataset that passed the *P* value threshold corresponding to FDR 5% in our initial mapping, and filtered out meQTLs that were in LD ($R^2 > 0.8$) keeping one SNP per independent loci (56,574 independent SNPs). For the resampling set, we considered all SNPs that were initially used for the meQTL mapping and pruned them for LD ($R^2 > 0.8$), yielding a final set of 921,466 SNPs. Resampling was performed using bins of allelic frequencies at intervals of 5%.

Finally, we tested for fold enrichments of meQTLs in GWAS hits, for each of the 17 parental EFO categories [75]. The fold enrichment was calculated by comparing the number of LD pruned-meQTLs that were found to correspond to GWAS hits (or were in LD with GWAS hits) with the expected number estimated through 10,000 resamples. *P* values associated to the fold enrichment were calculated by fitting a normal distribution to the empirical

distribution of our 10,000 resampled sets of SNPs. Confidence intervals were computed using 10,000 resamples by bootstrap. The same procedure was applied when searching for enrichments of meQTLs specifically in GWAS hits related to the 268 traits of the "Immune system disorder" EFO parental term.

Expression quantitative trait methylation (eQTM) analysis

To identify associations between DNA methylation levels and gene expression of nearby genes, we leveraged RNA-sequencing data obtained from the same individuals, both at the non-stimulated state (NS) and in response to four immune stimuli [48]. Briefly, RNA-sequencing was performed on the Illumina HiSeq2000 platform with 101-bp single-read sequencing with fragment size of around 295 bp, and outputs of around 30 million single-end reads per sample were obtained. A total of 763 RNA-sequencing samples from our filtered dataset of 156 donors were analyzed for gene expression profiling, including 156, 151, 153, 148, and 155 samples for the NS, LPS, Pam3CSK4, R848, and IAV conditions, respectively. Details of cell culture, immune stimulation conditions, and RNA-seq processing can be found in ref. [48].

Using the RNA-sequencing data from the NS condition, we mapped eQTMs (i.e., CpGs whose variation is associated with gene expression) in a window of 100 kb around the TSS of each gene (12,578 expressed genes in primary monocytes). The associated *P* values and the coefficients of correlation between methylation profiles and gene expression were obtained using Spearman's rank correlation. FDR was computed by mapping eQTMs on 100 datasets with the *M* values permuted, and kept, after each permutation, the most significant *P* value per gene (gene-level FDR). We selected the *P* value threshold that provided a 5% FDR ($P = 5 \times 10^{-5}$).

We also mapped eQTMs in the context of the response to the various stimulations, namely response-QTMs (reQTMs). To do so, the same procedure explained above for the eQTM mapping was followed, using the fold change of expression upon stimulation as a measure of the host response to infection. Specifically, we calculated the difference of the $\log_2$ of expression values between the stimulated and non-stimulated states, corrected for the effect of low values of FPKM, for each gene expressed in at least one of the two conditions.

$$\begin{aligned} Diff &= \log_2(1 + FPKM_{Stim}) - \log_2(1 + FPKM_{NS}) \\ &= \log_2\left(\frac{1 + FPKM_{Stim}}{1 + FPKM_{NS}}\right) \end{aligned}$$

$$FoldChange = \frac{1 + FPKM_{Stim}}{1 + FPKM_{NS}} = 2^{Diff}$$

For the mapping of eQTMs and reQTMs, we conducted two separate analyses: merging all individuals and including ancestry as a covariate, or keeping the two populations separately.

Expression quantitative trait loci (eQTL) analysis

We mapped expression quantitative trait loci (eQTLs) using the MatrixEQTL R package [70], leveraging our genotyping and expression data [48]. As for the meQTL mapping, we filtered out SNPs with a minor allele frequency < 5% in the populations studied and kept a final dataset of 10,278,745 SNPs. Age and PC1/PC2 of the genotype matrix were used as covariates in the linear model. Two different models were used: merging all individuals and including ancestry as a covariate, or keeping the two populations separately. We also mapped response quantitative trait loci (reQTLs), using the fold change of expression described above, instead of expression, and the same covariates that we used for the eQTL mapping.

For both eQTLs and reQTLs, FDR was computed by mapping eQTLs/reQTLs on 100 datasets with the expression values permuted within each population. We then kept, after each permutation, the most significant *P* value per gene, across populations (gene-level FDR). Finally, we computed the FDR associated with different *P* value thresholds for eQTLs or reQTLs, and subsequently selected the *P* value threshold that provided a 5% FDR: $P = 5 \times 10^{-5}$ and $P = 5 \times 10^{-6}$ for eQTLs and reQTLs, respectively.

Simulations to infer causality

We simulated different scenarios to infer causal relationships between DNA methylation and gene expression. For each scenario, we started by randomly selecting genomic blocks of 1 Mb each along the genome to keep realistic expectations of genetic structure. We next randomly sampled SNPs in these blocks, which we used to simulate methylation and gene expression data. For example, in a scenario where a genetic variant influences DNA methylation variation that, in turn, actively regulates gene expression (see Fig. 4a), we followed the next steps:

(i)

$$G_{i\_std} = \frac{(G_i - \overline{G_i})}{sd\ (G_i)}$$

(ii)

$$M_i = \sqrt{\alpha_i} \times G_{i_{std}} + \sqrt{(1-\alpha_i)} \times \varepsilon_i$$

(iii)

$$M_{i\_std} = \frac{(M_i - \overline{M_i})}{sd\ (M_i)}$$

(iv)

$$E_i = \sqrt{\gamma \times \beta_i} \times M_{i_{std}} + \sqrt{\gamma \times \tau_i} \times G_{i_{std}} + \sqrt{\left(1-\gamma \times (\beta_i + \tau_i)\right)} \times \zeta_i$$

where G_i is the genotype of the ith sampled variant and G_{i_std} the standardized value of its genotype; M_i is the simulated methylation data and M_{i_std} its standardized methylation value; E_i is the simulated gene expression data; α_i is the proportion of variance of M_i that is explained by G_i, and γ is a noise parameter that corresponds to the total proportion of variance of E_i that is explained by G_i and M_i. β_i and τ_i are the proportions of explained variance that are attributable to G_i and M_i respectively (satisfying $\beta_i + \tau_i = 1$). Finally, ε_i and ζ_i are random, normally distributed residuals. Note that in the simulation presented in Fig. 4a, b, we used a gamma of 0.25, so that 75% of the variance of gene expression remained unexplained.

Detection of genetic variants-DNA methylation-gene expression trios

To infer causality between regulatory loci and gene expression variation, we considered eQTLs that were also detected as meQTLs, and, out of this subset, we kept only those for which the meQTL-CpG had previously been identified as an eQTM of the eQTL-gene (Additional file 1: Figure S11). When multiple SNPs or CpGs where present in a trio, we used an elastic net model, to build linear predictors of (i) gene expression based on DNA methylation variability for trios with multiple CpGs and (ii) DNA methylation based on genetic variability for trios with multiple SNPs. These predictors were then used as summary variables for DNA methylation variability (i) or genetic variability (ii). Specifically, the *glmnet* function from the R package glmnet [106] was used to fit the generalized linear model via penalized maximum likelihood, with an elastic net mixing parameter α of 0.5. The strength of the penalty λ_{1se} was chosen as the largest value of lambda such that the error was within 1 standard deviation of the minimum lambda, when performing k-fold cross validation with the *cv.glmnet* function. Finally, the generic R function *predict* was used to build the optimal linear predictor in each case. For the trios presenting more than one SNP, we also used a predictor of gene expression based on genetic variability, as summary variable for the genetic variability, and found no differences in our simulation-based mediation results when compared to building the summary variable from a predictor of DNA methylation (data not shown).

Mediation analyses

For conducting causal mediation analyses, we used a Bayesian approach as implemented in the mediation R package [90]. Briefly, this approach estimates causal effects of a mediating variable M (DNA methylation) on the relationship between an independent variable X (genetics) and a dependent variable Y (gene expression). In this scenario, the global effect of X on Y can be written as $\rho_{X \to Y} = \tau + \alpha \cdot \beta$, where τ is the specific effect of X on Y, α the specific effect of X on M, and β the specific effect of M on Y. With this, the product $\alpha \cdot \beta$ represents the mediation effect of G on Y, through M. The *mediate* function of the mediation R package was used to compute point estimates for average causal mediation effects, as well as 1000 simulation draws of average causal mediation effects. The empirical distribution of simulated effects was used to fit a normal distribution, which was subsequently used to compute empirical P values for the H_0 hypothesis "$\alpha \cdot \beta = 0$." We used the R function *p.adjust* with method "fdr" to correct at a FDR = 5%.

For comparison purposes with the mediation analyses, we conducted on simulated data a partial correlation approach to test for independence between expression and methylation levels when accounting for genetic variability. We used the *pcor.test* function from the R package ppcor [107] to compute P values of the partial correlation between simulated expression and methylation data.

Additional files

Additional file 1: Figure S1. Overview of the EvoImmunoPop experimental setting. **Figure S2.** Exploring the non-linear effects of age on DNA methylation. **Figure S3.** Mono-DMS were detected using the same approach as described in the Materials and methods section, and including the proportions in monocyte subpopulations as covariates. **Figure S4.** PCA of the genetic data, based on 151,419 SNPs, for Africans (AFB, red dots) and Europeans (EUB, blue dots). **Figure S5.** Fine mapping of meQTLs. **Figure S6.** Histogram of physical proximity of *cis*-meQTLs. **Figure S7.** Genomic location of CpG sites associated with a meQTL. **Figure S8.** Proportions of population differences in DNA methylation accounted for by genetics. **Figure S9.** Fold enrichment of meQTLs in GWAS hits. **Figure S10.** Fold enrichment of meQTLs associated with DMS in GWAS hits related to "immune system disorder". **Figure S11.** Rationale for the detection of trios to be used for causality inference. **Figure S12.** Cartoons of the various simulated scenarios. **Figure S13.** Heat map of correlation between the first ten PCs of expression and DNA methylation. **Figure S14.** Genomic location of eQTMs (NS) and reQTMs (for all stimulated conditions). **Figure S15.** Number of reQTM-genes, per condition, according to the direction of their association with DNA methylation. **Figure S16.** Causality inference upon immune stimulation. (PDF 3088 kb)

Additional file 2: Notes 1–2. (PDF 94 kb)

Additional file 3: Table S1. (XLSX 868 kb)

Additional file 4: Table S2. (XLSX 75 kb)

Additional file 5: Table S3. (XLSX 157 kb)

Additional file 6: Table S4. (XLSX 36 kb)

Funding

This project was funded by the Institut Pasteur, the CNRS, and the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC grant agreement 281297 (to L.Q.-M.). M.R. was supported by a Marie Skłodowska-Curie fellowship (DLV-655417).

Availability of data and materials

The DNA methylation data generated in this study have been deposited in the NCBI Gene Expression Omnibus (GEO) under accession code GSE120610 [108]. Genome-wide SNP genotyping, whole exome sequencing, and RNA-sequencing data used in this study are available at the European Genome-Phenome Archive (EGA) under accession code EGAS00001001895 [109].

Authors' contributions

LTH designed and performed the computational analyses, analyzed the data, and interpreted the results, with input from MR, MF, HQ, HA, EP, and LQ-M. LMM, JLM, and MSK contributed the DNA methylation data. NZ contributed the flow cytometry data. MR, HA, and EP contributed with ideas and participated in evaluating results and discussions. LQ-M conceived and supervised the study and obtained the funding. LTH and LQ-M wrote the manuscript, with input from all authors. All authors approved the final manuscript.

Ethics approval and consent to participate

All healthy donors provided informed consent. All experiments were approved by the Ethics Board of Institut Pasteur (EVOIMMUNOPOP-281297) and the relevant French authorities (CPP, CCITRS and CNIL), subject to applicable laws and regulations and ethical principles consistent with the Declaration of Helsinki.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France. [2]Centre National de la Recherche Scientifique (CNRS) UMR2000, 75015 Paris, France. [3]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, 75015 Paris, France. [4]Laboratory for Epigenetics & Environment, Centre National de Recherche en Génomique Humaine (CNRGH), CEA-Institut de Biologie François Jacob, 91000 Evry, France. [5]Department of Medical Genetics, University of British Columbia, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, Vancouver, BC, Canada.

Received: 14 June 2018 Accepted: 4 December 2018
Published online: 18 December 2018

References

1. Brinkworth JF, Barreiro LB. The contribution of natural selection to present-day susceptibility to chronic inflammatory and autoimmune disease. Curr Opin Immunol. 2014;31:66–78.
2. Casanova JL, Abel L, Quintana-Murci L. Immunology taught by human genetics. Cold Spring Harb Symp Quant Biol. 2013;78:157–72.
3. Casanova JL. Severe infectious diseases of childhood as monogenic inborn errors of immunity. Proc Natl Acad Sci U S A. 2015;112:E7128–37.
4. Casanova JL. Human genetic basis of interindividual variability in the course of infection. Proc Natl Acad Sci U S A. 2015;112:E7118–27.
5. Fumagalli M, Sironi M. Human genome variability, natural selection and infectious diseases. Curr Opin Immunol. 2014;30:9–16.
6. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011;7:e1002355.
7. Casanova JL, Abel L. Inborn errors of immunity to infection: the rule rather than the exception. J Exp Med. 2005;202:197–201.
8. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014;15:379–93.
9. Quintana-Murci L, Clark AG. Population genetic tools for dissecting innate immunity in humans. Nat Rev Immunol. 2013;13:280–93.
10. Siddle KJ, Quintana-Murci L. The Red Queen's long race: human adaptation to pathogen pressure. Curr Opin Genet Dev. 2014;29:31–8.
11. Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet. 2010;11:17–30.
12. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14:204–20.
13. Schubeler D. Function and information content of DNA methylation. Nature. 2015;517:321–6.
14. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2011;13:97–109.
15. Kaminsky ZA, Tang T, Wang SC, Ptak C, Oh GH, Wong AH, Feldcamp LA, Virtanen C, Halfvarson J, Tysk C, et al. DNA methylation profiles in monozygotic and dizygotic twins. Nature Genet. 2009;41:240–5.
16. Lam LL, Emberly E, Fraser HB, Neumann SM, Chen E, Miller GE, Kobor MS. Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci U S A. 2012;109(Suppl 2):17253–60.
17. Ziller MJ, Gu H, Muller F, Donaghey J, Tsai LT, Kohlbacher O, De Jager PL, Rosen ED, Bennett DA, Bernstein BE, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.
18. Marr AK, MacIsaac JL, Jiang R, Airo AM, Kobor MS, McMaster WR. Leishmania donovani infection causes distinct epigenetic DNA methylation changes in host macrophages. PLoS Pathog. 2014;10:e1004419.
19. Pacis A, Tailleux L, Morin AM, Lambourne J, MacIsaac JL, Yotova V, Dumaine A, Danckaert A, Luca F, Grenier JC, et al. Bacterial infection remodels the DNA methylation landscape of human dendritic cells. Genome Res. 2015;25:1801–11.
20. Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai SL, Arepalli S, Dillman A, Rafferty IP, Troncoso J, et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLoS Genet. 2010;6:e1000952.
21. Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu C. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet. 2010;86:411–9.
22. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. Genome Biol. 2011;12:R10.
23. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. eLife. 2013;2:e00523.
24. Banovich NE, Lan X, McVicker G, van de Geijn B, Degner JF, Blischak JD, Roux J, Pritchard JK, Gilad Y. Methylation QTLs are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. PLoS Genet. 2014;10:e1004663.
25. Olsson AH, Volkov P, Bacos K, Dayeh T, Hall E, Nilsson EA, Ladenvall C, Ronn T, Ling C. Genome-wide associations between genetic and epigenetic variation influence mRNA expression and insulin secretion in human pancreatic islets. PLoS Genet. 2014;10:e1004735.
26. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, St Clair D, Mustard C, Breen G, Therman S, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biol. 2016;17:176.
27. Hannon E, Spiers H, Viana J, Pidsley R, Burrage J, Murphy TM, Troakes C, Turecki G, O'Donovan MC, Schalkwyk LC, et al. Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. Nat Neurosci. 2016;19:48–54.
28. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014;15:R37.
29. van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, Ehli EA, Davies GE, van Iterson M, Breeze CE, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016;7:11115.
30. McClay JL, Shabalin AA, Dozmorov MG, Adkins DE, Kumar G, Nerella S, Clark SL, Bergen SE, Swedish Schizophrenia C, Hultman CM, et al. High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction. Genome Biol. 2015;16:291.
31. Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A, et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 2012;8:e1002629.
32. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC, Ried JS, Zhang W, Yang Y, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6.

33. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70:27–56.
34. Baylin SB, Jones PA. Epigenetic determinants of cancer. Cold Spring Harb Perspect Biol. 2016;8(9):a019505.
35. Bell CG, Finer S, Lindgren CM, Wilson GA, Rakyan VK, Teschendorff AE, Akan P, Stupka E, Down TA, Prokopenko I, et al. Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. PLoS One. 2010;5:e14040.
36. Chambers JC, Loh M, Lehne B, Drong A, Kriebel J, Motta V, Wahl S, Elliott HR, Rota F, Scott WR, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. Lancet Diabetes Endocrinol. 2015;3:526–34.
37. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol. 2013;31:142–7.
38. Heyn H, Moran S, Hernando-Herraez I, Sayols S, Gomez A, Sandoval J, Monk D, Hata K, Marques-Bonet T, Wang L, Esteller M. DNA methylation contributes to natural human variation. Genome Res. 2013;23:1363–72.
39. Moen EL, Zhang X, Mu W, Delaney SM, Wing C, McQuade J, Myers J, Godley LA, Dolan ME, Zhang W. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. Genetics. 2013;194:987–96.
40. Fraser HB, Lam LL, Neumann SM, Kobor MS. Population-specificity of human DNA methylation. Genome Biol. 2012;13:R8.
41. Fagny M, Patin E, MacIsaac JL, Rotival M, Flutre T, Jones MJ, Siddle KJ, Quach H, Harmant C, McEwen LM, et al. The epigenomic landscape of African rainforest hunter-gatherers and farmers. Nat Commun. 2015;6:10047.
42. Carja O, MacIsaac JL, Mah SM, Henn BM, Kobor MS, Feldman MW, Fraser HB. Worldwide patterns of human epigenetic variation. Nat Ecol Evol. 2017;1:1577–83.
43. Galanter JM, Gignoux CR, Oh SS, Torgerson D, Pino-Yanes M, Thakur N, Eng C, Hu D, Huntsman S, Farber HJ, et al. Differential methylation between ethnic sub-groups reflects the effect of genetic ancestry and environmental exposures. Elife. 2017;6:e20532.
44. Gopalan S, Carja O, Fagny M, Patin E, Myrick JW, McEwen LM, Mah SM, Kobor MS, Froment A, Feldman MW, et al. Trends in DNA methylation with age replicate across diverse human populations. Genetics. 2017;206:1659–74.
45. Sugawara H, Iwamoto K, Bundo M, Ueda J, Ishigooka J, Kato T. Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines. Epigenetics. 2011;6:508–15.
46. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
47. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nat Rev Genet. 2018;19:129–47.
48. Quach H, Rotival M, Pothlichet J, Loh YE, Dannemann M, Zidane N, Laval G, Patin E, Harmant C, Lopez M, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. Cell. 2016;167:643–56 e617.
49. Nedelec Y, Sanz J, Baharian G, Szpiech ZA, Pacis A, Dumaine A, Grenier JC, Freiman A, Sams AJ, Hebert S, et al. Genetic ancestry and natural selection drive population differences in immune responses to pathogens. Cell. 2016;167:657–69 e621.
50. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012;109:1204–9.
51. Fairfax BP, Humburg P, Makino S, Naranbhai V, Wong D, Lau E, Jostins L, Plant K, Andrews R, McGee C, Knight JC. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science. 2014;343:1246949.
52. Lee MN, Ye C, Villani AC, Raj T, Li W, Eisenhaure TM, Imboywa SH, Chipendo PI, Ran FA, Slowikowski K, et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science. 2014;343:1246980.
53. Caliskan M, Baker SW, Gilad Y, Ober C. Host genetic variation influences gene expression response to rhinovirus infection. PLoS Genet. 2015;11:e1005111.
54. Kim S, Becker J, Bechheim M, Kaiser V, Noursadeghi M, Fricker N, Beier E, Klaschik S, Boor P, Hess T, et al. Characterizing the genetic basis of innate immune response in TLR4-activated human monocytes. Nat Commun. 2014;5:5236.
55. Kim-Hellmuth S, Bechheim M, Putz B, Mohammadi P, Nedelec Y, Giangreco N, Becker J, Kaiser V, Fricker N, Beier E, et al. Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. Nat Commun. 2017;8:266.
56. Pai AA, Pritchard JK, Gilad Y. The genetic and mechanistic basis for variation in gene regulation. PLoS Genet. 2015;11:e1004857.
57. Yang X, Han H, De Carvalho DD, Lay FD, Jones PA, Liang G. Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell. 2014;26:577–90.
58. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. Oncotarget. 2012;3:462–74.
59. Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y, et al. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature. 2010;466:253–7.
60. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, Bryois J, Padioleau I, Romano L, Planchon A, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. PLoS Genet. 2015;11:e1004958.
61. Relton CL, Davey Smith G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. Int J Epidemiol. 2012;41:161–76.
62. Richardson TG, Zheng J, Davey Smith G, Timpson NJ, Gaunt TR, Relton CL, Hemani G. Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. Am J Hum Genet. 2017;101:590–602.
63. Hannon E, Weedon M, Bray N, O'Donovan M, Mill J. Pleiotropic effects of trait-associated genetic variation on DNA methylation: utility for refining GWAS loci. Am J Hum Genet. 2017;100:954–9.
64. Bell CG, Gao F, Yuan W, Roos L, Acton RJ, Xia Y, Bell J, Ward K, Mangino M, Hysi PG, et al. Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci. Nat Commun. 2018;9:8.
65. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, van Iterson M, van Dijk F, van Galen M, Bot J, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49:131–8.
66. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94:559–73.
67. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22:1748–59.
68. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
69. Johnson ND, Wiener HW, Smith AK, Nishitani S, Absher DM, Arnett DK, Aslibekyan S, Conneely KN. Non-linear patterns in age-related DNA methylation may reflect CD4(+) T cell differentiation. Epigenetics. 2017;12:492–503.
70. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28:1353–8.
71. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, et al. Genetics of gene expression and its effect on disease. Nature. 2008;452:423–8.
72. Dermitzakis ET. Cellular genomics for complex traits. Nat Rev Genet. 2012;13:215–20.
73. Nica AC, Montgomery SB, Dimas AS, Stranger BE, Beazley C, Barroso I, Dermitzakis ET. Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. PLoS Genet. 2010;6:e1000895.
74. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009;10:184–94.
75. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 2010;26:1112–8.
76. Gu J, Stevens M, Xing X, Li D, Zhang B, Payton JE, Oltz EM, Jarvis JN, Jiang K, Cicero T, et al. Mapping of variable DNA methylation across multiple cell types defines a dynamic regulatory landscape of the human genome. G3 (Bethesda). 2016;6:973–86.
77. Hannon E, Lunnon K, Schalkwyk L, Mill J. Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic

studies of neurological and neuropsychiatric phenotypes. Epigenetics. 2015; 10:1024–32.
78. Cheung WA, Shao X, Morin A, Siroux V, Kwan T, Ge B, Aissi D, Chen L, Vasquez L, Allum F, et al. Functional variation in allelic methylomes underscores a strong genetic contribution and reveals novel epigenetic alterations in the human epigenome. Genome Biol. 2017;18:50.
79. Schultz MD, He Y, Whitaker JW, Hariharan M, Mukamel EA, Leung D, Rajagopal N, Nery JR, Urich MA, Chen H, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523:212–6.
80. Farre P, Jones MJ, Meaney MJ, Emberly E, Turecki G, Kobor MS. Concordant and discordant DNA methylation signatures of aging in human blood and brain. Epigenetics Chromatin. 2015;8:19.
81. Pai AA, Bell JT, Marioni JC, Pritchard JK, Gilad Y. A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues. PLoS Genet. 2011;7:e1001316.
82. Nititham J, Taylor KE, Gupta R, Chen H, Ahn R, Liu J, Seielstad M, Ma A, Bowcock AM, Criswell LA, et al. Meta-analysis of the TNFAIP3 region in psoriasis reveals a risk haplotype that is distinct from other autoimmune diseases. Genes Immun. 2015;16:120–6.
83. Yin X, Low HQ, Wang L, Li Y, Ellinghaus E, Han J, Estivill X, Sun L, Zuo X, Shen C, et al. Genome-wide meta-analysis identifies multiple novel associations and ethnic heterogeneity of psoriasis susceptibility. Nat Commun. 2015;6:6916.
84. Bonder MJ, Kasela S, Kals M, Tamm R, Lokk K, Barragan I, Buurman WA, Deelen P, Greve JW, Ivanov M, et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. BMC Genomics. 2014;15:860.
85. Ecker S, Chen L, Pancaldi V, Bagger FO, Fernandez JM, Carrillo de Santa Pau E, Juan D, Mann AL, Watt S, Casale FP, et al. Genome-wide analysis of differential transcriptional and epigenetic variability across human immune cell types. Genome Biol. 2017;18:18.
86. Shlyueva D, Stampfel G, Stark A. Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet. 2014;15:272–86.
87. Rickels R, Shilatifard A. Enhancer logic and mechanics in development and disease. Trends Cell Biol. 2018;28(8):608–30.
88. Yao L, Berman BP, Farnham PJ. Demystifying the secret mission of enhancers: linking distal regulatory elements to target genes. Crit Rev Biochem Mol Biol. 2015;50:550–73.
89. Grundberg E, Meduri E, Sandling JK, Hedman AK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M, 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. 2013;93:876–90.
90. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R package for causal mediation analysis. J Stat Softw. 2014;59:1–38.
91. MacKinnon DP. Multivariate applications series. Introduction to statistical mediation analysis. New York: Taylor & Francis Group/Lawrence Erlbaum Associates; 2008.
92. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011;480:490–5.
93. Stroud H, Feng S, Morey Kinney S, Pradhan S, Jacobsen SE. 5-Hydroxymethylcytosine is associated with enhancers and gene bodies in human embryonic stem cells. Genome Biol. 2011;12:R54.
94. Choi I, Kim R, Lim HW, Kaestner KH, Won KJ. 5-Hydroxymethylcytosine represses the activity of enhancers in embryonic stem cells: a new epigenetic signature for gene regulation. BMC Genomics. 2014;15:670.
95. Song MA, Brasky TM, Marian C, Weng DY, Taslim C, Dumitrescu RG, Llanos AA, Freudenheim JL, Shields PG. Racial differences in genome-wide methylation profiling and gene expression in breast tissues from healthy women. Epigenetics. 2015;10:1177–87.
96. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011;25:2227–41.
97. Valeri L, Reese SL, Zhao S, Page CM, Nystad W, Coull BA, London SJ. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? Epigenomics. 2017; 9:253–65.
98. Wang F, Zhang S, Wen Y, Wei Y, Yan H, Liu H, Su J, Zhang Y, Che J. Revealing the architecture of genetic and epigenetic regulation: a maximum likelihood model. Brief Bioinform. 2014;15:1028–43.
99. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015;526:68–74.
100. Du P, Kibbe WA, Lin SM. lumi: a pipeline for processing Illumina microarray. Bioinformatics. 2008;24:1547–8.
101. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
102. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29:189–96.
103. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
104. Smyth GK. Limma: linear models for microarray data. In: Gentleman R, Carey V, Dudoit S, Irizarry I, Hube W, editors. Bioinformatics and computational biology solutions using R and Bioconductor. New York: Springer; 2005. p. 397–420.
105. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45:D896–901.
106. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33:1–22.
107. Kim S. ppcor: an R package for a fast calculation to semi-partial correlation coefficients. Commun Stat Appl Methods. 2015;22:665–74.
108. Husquin LT, Rotival M, Fagny M, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. GEO. 2018; https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE120610. Accessed 30 Oct 2018.
109. Quach H, Rotival M, Pothlichet J, et al. Genetic adaptation and Neandertal admixture shaped the immune system of human populations. EGA. 2016; https://www.ebi.ac.uk/ega/studies/EGAS00001001895. Accessed 20 Oct 2016.
110. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.
111. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nat Protoc. 2017;12:2478–92.

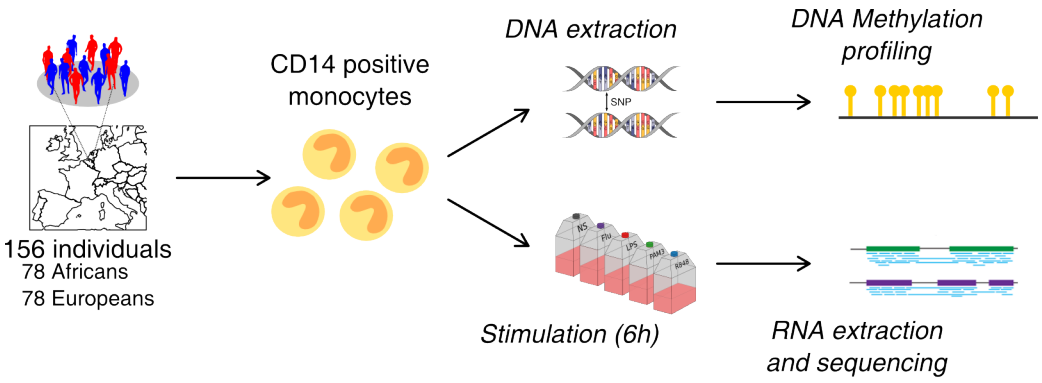

Figure S1 Overview of the EvoImmunoPop experimental setting. DNA methylation profiles and transcriptional responses to various immune stimulations, of primary monocytes from 156 healthy donors of European and African descent.

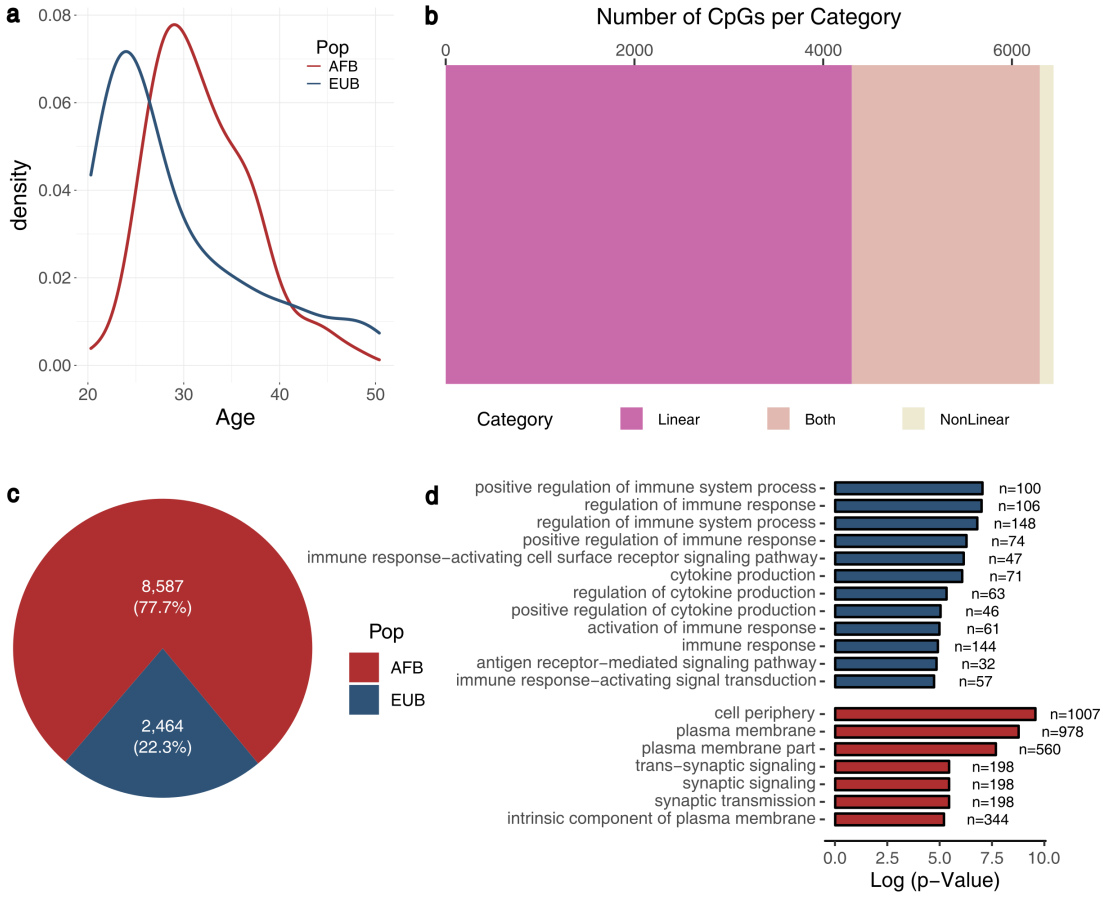

Figure S2 Exploring the non-linear effects of age on DNA methylation. **a** Age density in AFB (red) and EUB (blue) individuals. **b** Number of CpG sites presenting age-dependent methylation changes, as detected by linear regression (magenta), ANOVA (yellow) or both analyses (salmon). **c** Proportion of DMS detected with the ANOVA model that are either hypermethylated in AFB (red) or in EUB (blue) individuals. **d** Gene Ontology (GO) enrichment analyses of AFB- and EUB-DMS. For both groups, the top-GO categories reaching 5% FDR are shown, together with the number of genes per category and the log-transformed FDR-adjusted enrichment *P*-values.

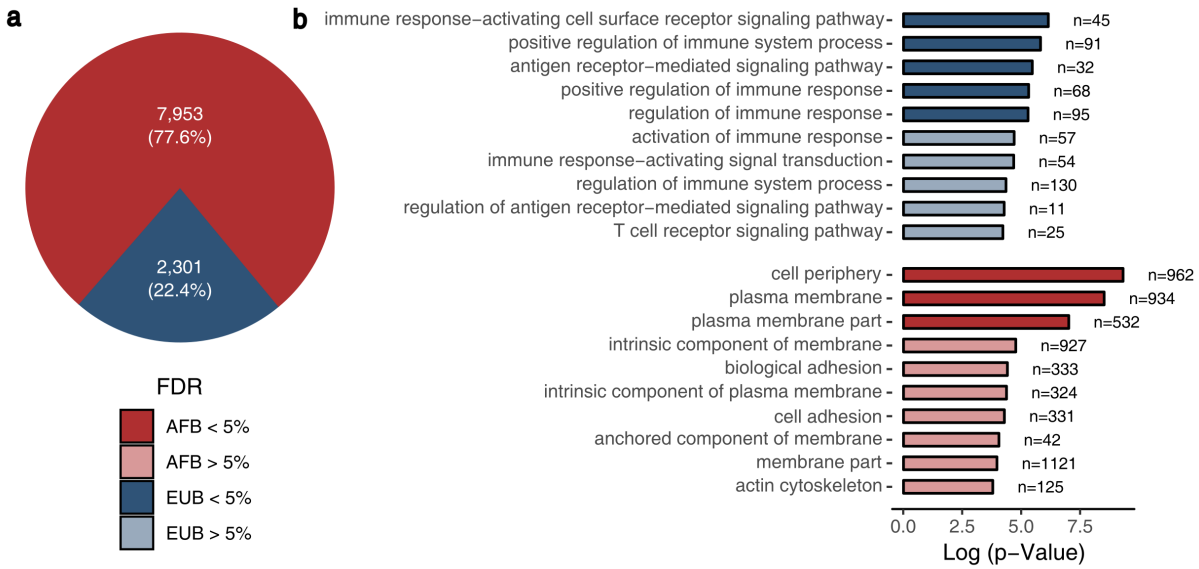

Figure S3 Mono-DMS were detected using the same approach as described in the Methods section, and including the proportions in monocyte subpopulations as covariates. **a** Proportion of Mono-DMS that are either hypermethylated in AFB (red) or in EUB (blue) individuals. **b** Gene Ontology (GO) enrichment analyses of AFB- and EUB-Mono-DMS. For both groups, the ten top-GO categories are shown. Categories reaching 5% FDR are shown in dark red for AFB and dark blue for EUB. The number of genes per category and the log-transformed FDR-adjusted enrichment *P*-values are also indicated.

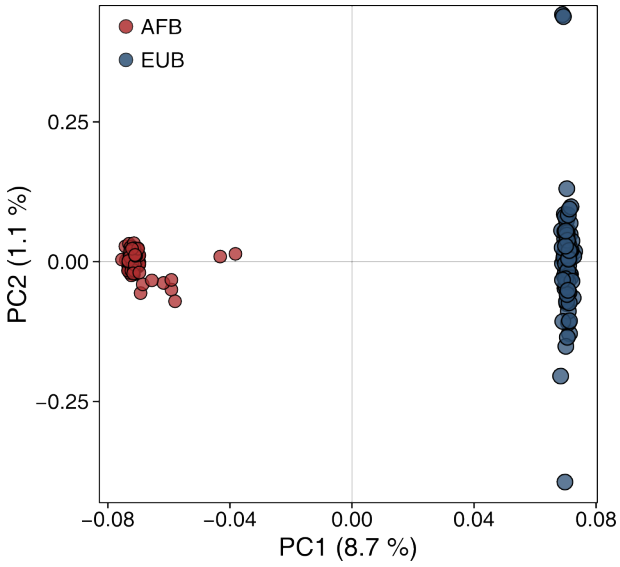

Figure S4 PCA of the genetic data, based on 151,419 SNPs, for Africans (AFB, red dots) and Europeans (EUB, blue dots). The percentages of variance explained by PC1 and PC2 are indicated.

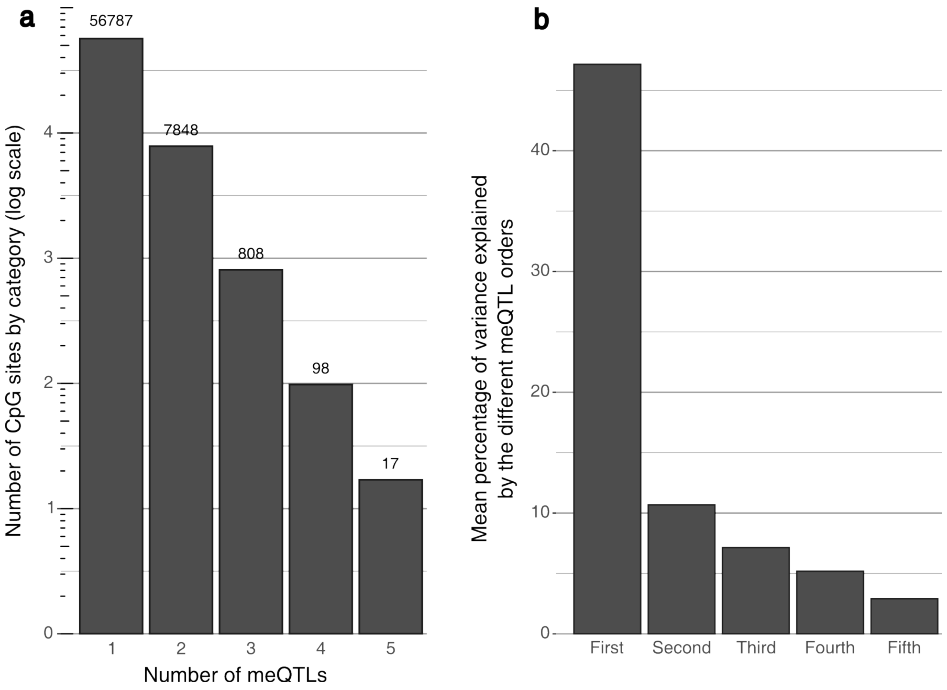

Figure S5 Fine mapping of meQTLs. **a** The number of CpG sites according to the number of associated independent meQTLs is shown on a log-scale. **b** Mean percentage of variance of DNA methylation explained by meQTLs according to the order in which they were detected.

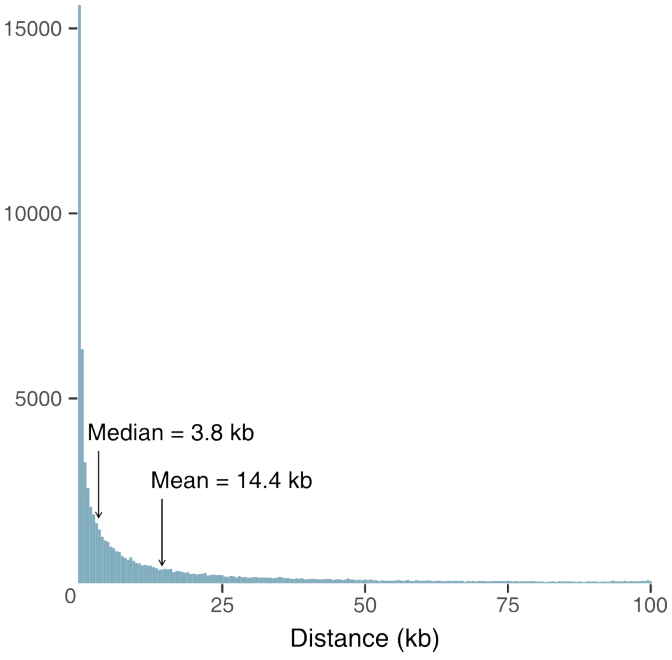

Figure S6 Histogram of physical proximity of *cis*-meQTLs. The distribution of the distances (in kb) between each meQTL and its associated CpG sites is presented, together with the mean and the median value.

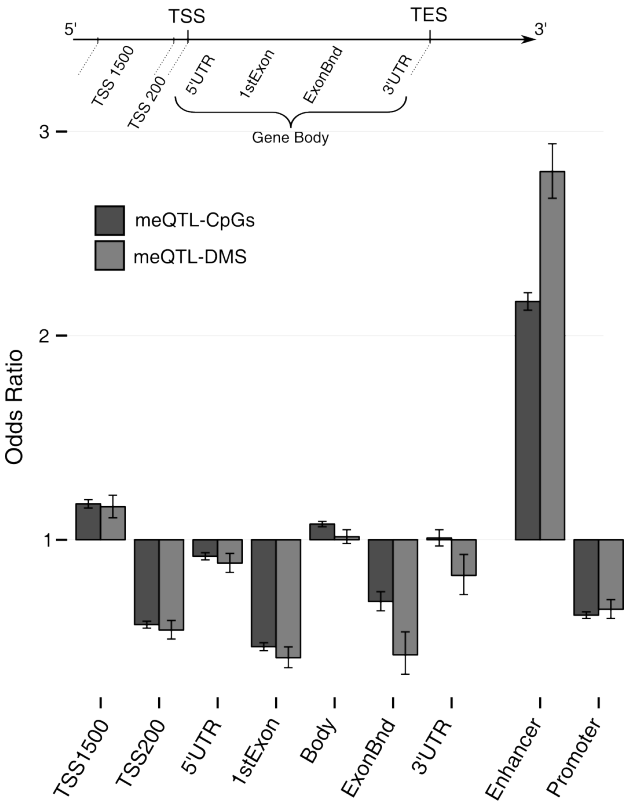

Figure S7 Genomic location of CpG sites associated with a meQTL. meQTL-CpGs are represented in dark grey, and the subset of these CpG sites that were also detected as DMS (meQTL-DMS) in light grey. OR were computed against the general distribution of the 552,141 CpGs of our dataset

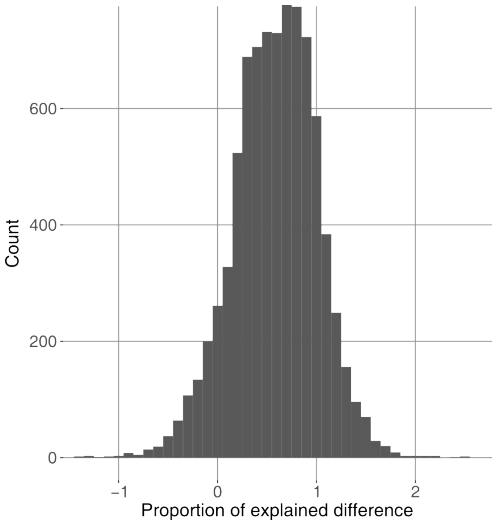

Figure S8 Proportions of population differences in DNA methylation accounted for by genetics. Histogram of the distribution of these proportions, for the 8,459 DMS that were associated with at least one meQTL. Proportions lower than 0 represent situations where genetics has an opposite effect to the observed overall population difference in DNA methylation. Conversely, proportions higher than 1 represent situations where the difference attributable to genetics is higher than that actually observed, indicative of an opposite effect of environmental factors or non-detected independent genetic effects.

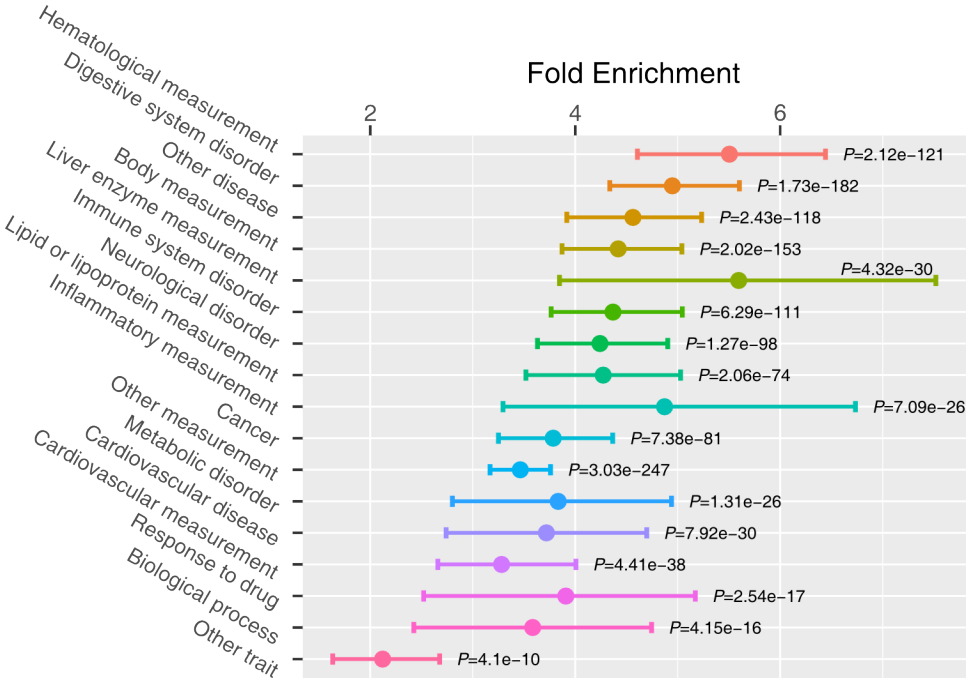

Figure S9 Fold enrichment of meQTLs in GWAS hits. For each of the 17 parental EFO categories, the fold enrichment, the 95% confidence intervals and the associated *P* values are shown.

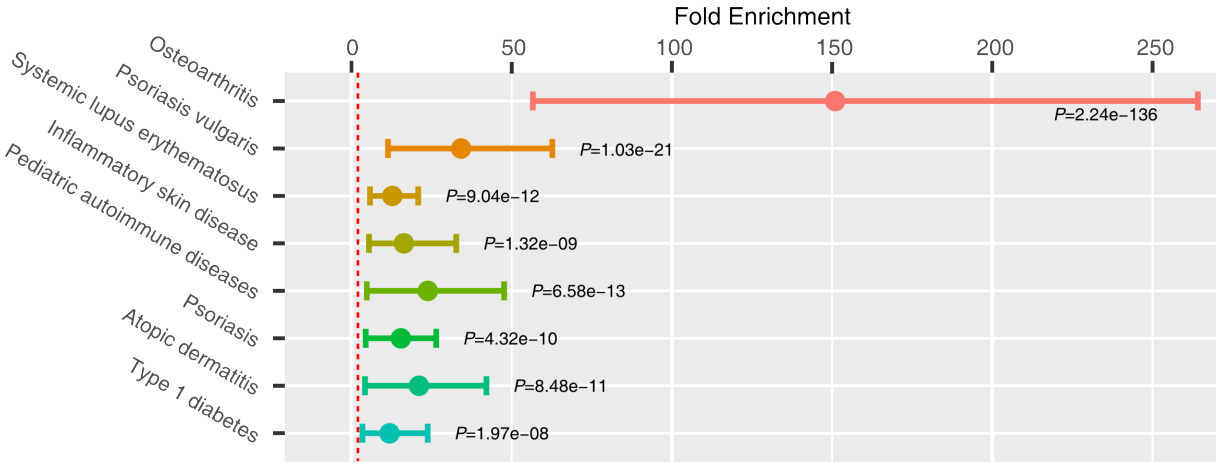

Figure S10 Fold enrichment of meQTLs associated with DMS in GWAS hits related to "immune system disorder". For the 8 signals that presented the higher lower-bound of confidence intervals, the fold enrichment, the 95% confidence intervals and the associated *P* values are shown.

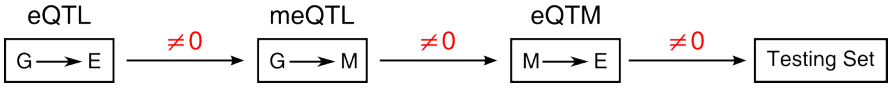

Figure S11 Rationale for the detection of trios to be used for causality inference.

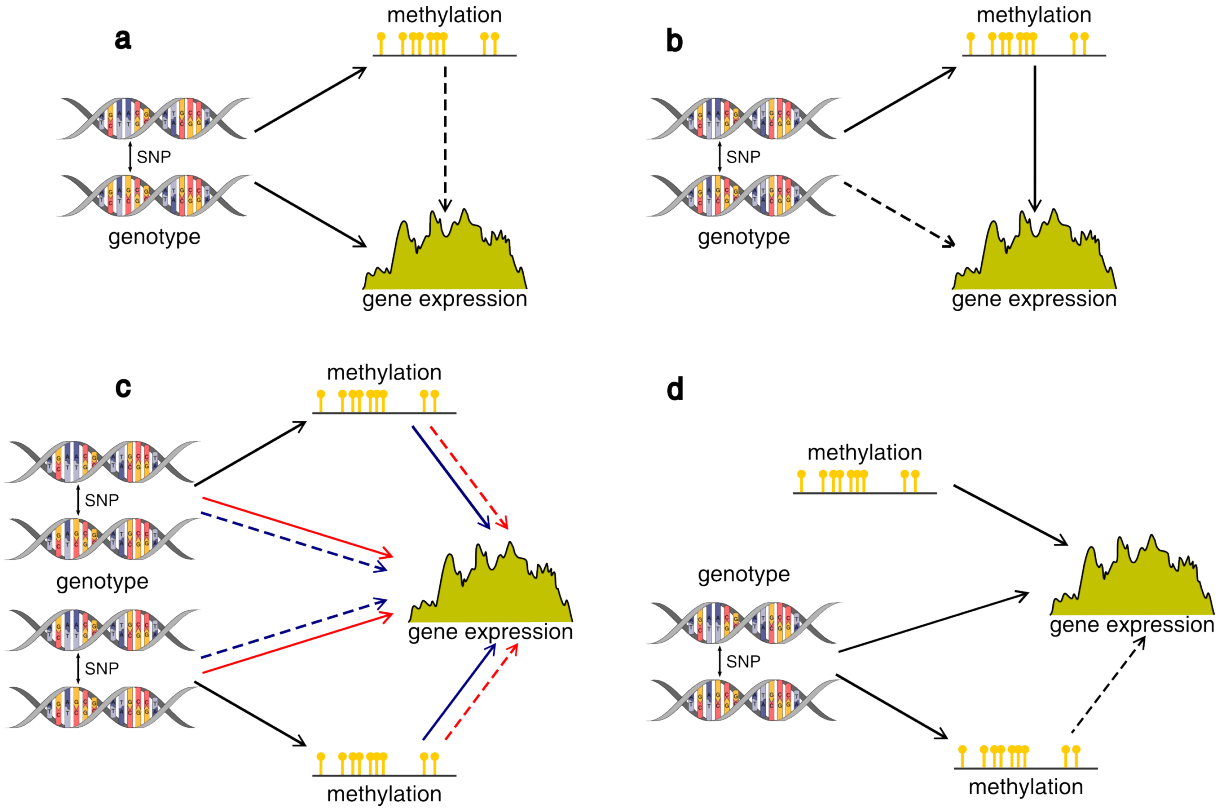

Figure S12 Cartoons of the various simulated scenarios. Plain arrows represent causal relationships, while dashed arrows represent correlations through another relationship. **a-b** Simple situations where either DNA methylation or genetics causally impact gene expression variation. **c** More complex scenarios where gene expression is causally impacted by two independent genetic (red arrows) or epigenetic (blue arrows) variants. **d** Scenario where the CpG site that causally impacts gene expression variation is not under the control of any genetic variant. Note that for all simulated scenarios (**a-d**), similar results between mediation analyses and partial correlations were obtained in terms of sensitivity and specificity (data not shown).

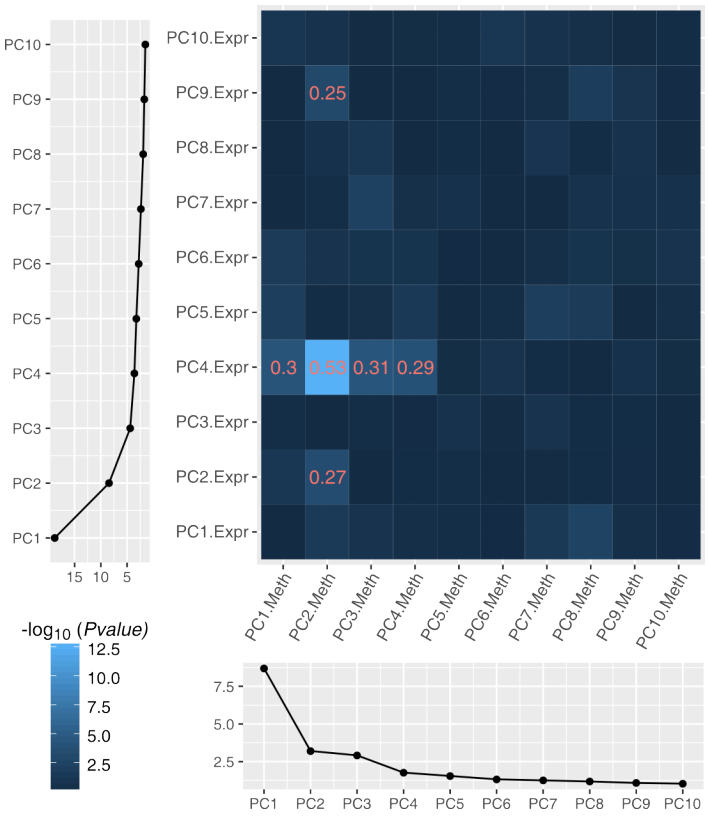

Figure S13 Heat map of correlation between the first ten PCs of expression and DNA methylation. Shades of blue are proportional with the –log10 of the correlation *P* values. In red are given the R^2 of the correlation for cases were $P < 0.001$. Bottom and left panels show the percentage of variance explained by the first ten PCs of gene expression and DNA methylation, respectively.

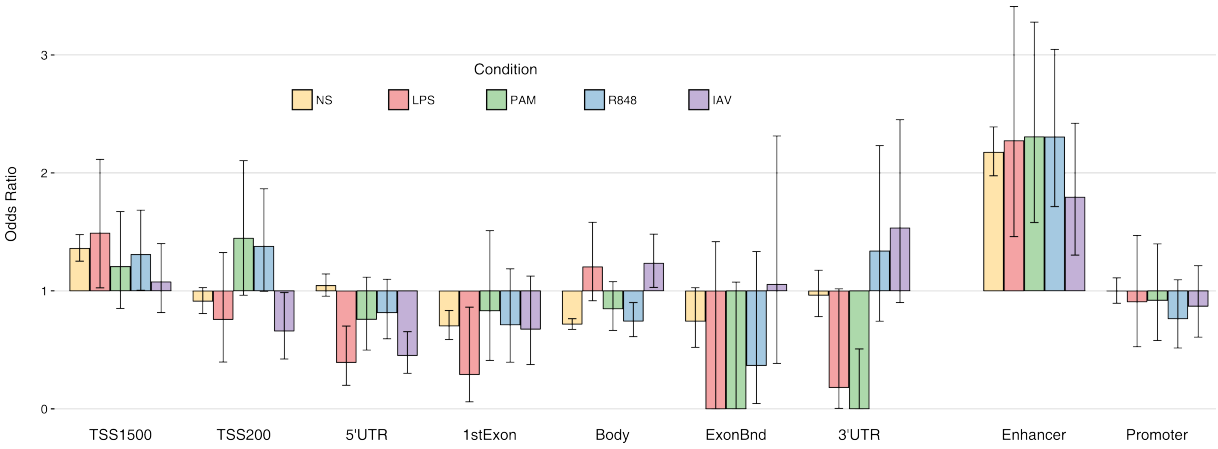

Figure S14 Genomic location of eQTMs (NS) and reQTMs (for all stimulated conditions). Odds ratio were computed against the general distribution of 552,141 CpGs of our dataset.

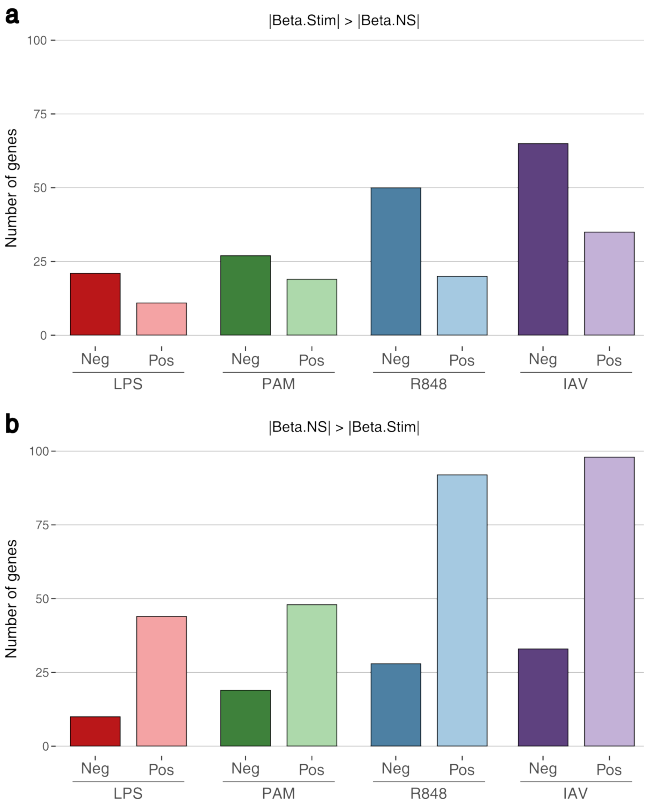

Figure S15 Number of reQTM-genes, per condition, according to the direction of their association with DNA methylation. **a** Cases presenting a stronger expression-methylation association upon stimulation than at the non-stimulated state, **b** Cases presenting a stronger expression-methylation association at the non-stimulated state than upon stimulation.

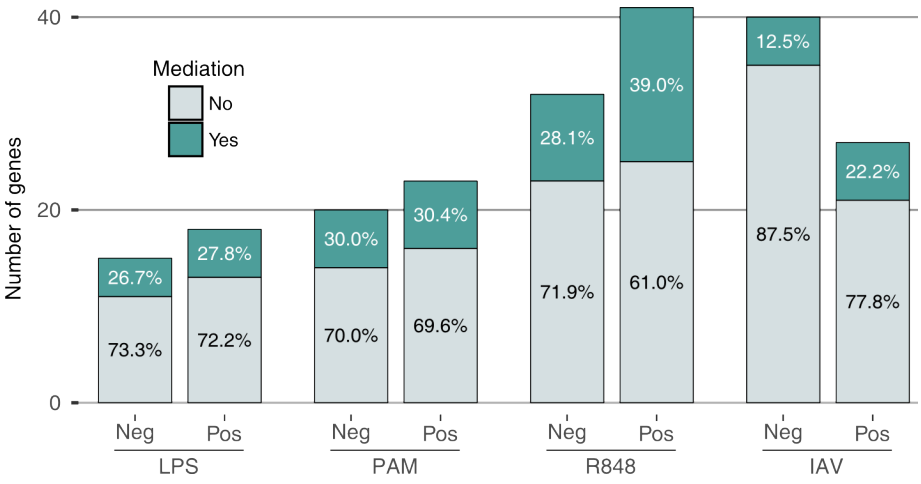

Figure S16 Causality inference upon immune stimulation. Number of mediated and non-mediated reQTM-genes for negative (Neg) and positive (Pos) associations between DNA methylation and fold-changes in expression upon different stimulation conditions. The percentages among these two categories are also indicated.

Note 1

Because the age distributions of African- and European-ancestry individuals significantly differ (Wilcoxon *P*-value = 10^{-4}; **Additional file 1: Fig. S2a**), we investigated the extent to which DNA methylation is non-linearly affected by age in our dataset.We first created a factor variable − grouping individuals by ranges of age (20-25, 25-30, 30-35, 35-40, 40-45, 45-50) − and subsequently applied an ANOVA regression of DNA methylation for each CpG site on this age factor variable, using the *Anova* function from the R package *car*.

$$Anova(lm(Meth \sim AgeClass + Pop)) \quad (i)$$

Concurrently, we aimed at identifying a linear effect of age on DNA methylation, using a linear model.

$$lm(Meth \sim Age + Pop) \quad (ii)$$

After correction for multiple testing using the *p.adjust* function from R, with method="fdr", we identified 6,295 and 2,137 CpGs sites that were detected by the linear and non-linear models, respectively. Interestingly, we found an important overlap between these two models, with 1,991 CpGs that were detected by both analyses (**Additional file 1: Fig. S2b**). This indicates that most of the effects of age on DNA methylation are captured by the linear model, with only 146 CpGs for which age has a strict non-linear effect. On the other hand, using a non-linear model appears to considerably reduce power, with 4,158 CpG sites that are found to be associated with age only when using the linear model. To ensure that the population differences in DNA methylation detected were not age-related effects that we failed to adjust for, we repeated our DMS mapping and accounted for putative non-linear effects of age, by using the *P*-values of the variable *Pop* in model (*i*). After Benjamini–Hochberg correction, we detected 50,353 CpGs (FDR = 1%) that presented a significant difference between AFB and EUB. When restricting this analysis to CpGs that presented a mean difference > 5%, we identified 11,051 DMS. The vast majority of these (11,001 DMS, 99.5%) were detected by our analysis that did not consider the non-linear effects of age. Furthermore, we found the same asymmetry in terms of number of DMS that are hypermethylated in AFB and EUB (**Additional file 1: Fig. S2c**), and similar GO enrichments (**Additional file 1: Fig. S2d**). Of the 1,049 DMS that were not detected by this analysis, with respect to the original 12,050 DMS, only 16 (1.3%) appear to be non-linearly affected by age. This suggests that the difference observed between analyses results from a difference in power, rather than a genuine non-linear effect of age that we failed to adjust for in our original DMS scan.

Note 2

The reverse causation scenario, where the impact of genetic variation on DNA methylation is mediated by gene expression variation, is highly unlikely in our experimental setting (**Additional file 1: Fig. S1**). Given that DNA methylation was obtained from monocytes at *t*=0, while gene expression was obtained at *t*=6h, the reverse causation could only be observed in cases where expression at *t*=6h is a proxy of expression at *t*=0. We nonetheless tested this hypothesis by considering three different models: *Model 1*, independent control of both gene expression and DNA methylation by genetics; *Model 2*, genetic control of DNA methylation mediated by gene expression; and *Model 3*, genetic control of gene expression mediated by DNA methylation. We computed the log-likelihood of these three models:

$$L(\text{Model 1}) = L(M|G) \times L(E|G)$$

$$L(\text{Model 2}) = L(M|E) \times L(E|G)$$

$$L(\text{Model 3}) = L(E|M) \times L(M|G)$$

with G being the genetic variant, M the CpG site, E the gene expression, and L(Y|X) the likelihood of the standard linear model, with Y as the dependent variable and X as the predictor.

We then calculated each model's probability using a uniform distribution of the priors.

$$(1) \qquad P(Model_i|Data) = \frac{P(Model_i) * P(Data|Model_i)}{\sum_i [P(Model_i) * P(Data|Model_i)]}$$

where *P* represents the probability of model i, and $P(Model_1)=P(Model_2)=P(Model_3)=1/3$. The equation (1) can then easily be simplified as:

$$(2) \qquad P(Model_i|Data) = \frac{Likelihood\ (Model_i)}{\sum_i [Likelihood\ (Model_i)]}$$

We calculated the probability of each model for all trios, and assigned each trio to the model presenting the highest probability, which we required to be higher than 0.9. If no models reached such a probability, the trio was declared non-significant. We found that reverse causation was indeed highly unlikely: at the non-stimulated state, only 3.1% of the trios were assigned to *Model 2*, while <1% of the trios were assigned to *Model 2* in the presence of immune stimulation.

5.3 Résumé des résultats et perspectives

Afin d'entrevoir les différences existant entre individus d'origine européenne et individus d'origine africaine, nous avons tout d'abord séparé les individus selon leurs profils de méthylation à l'aide d'une analyse en composante principale. Cette première analyse nous a permis de montrer une assez nette séparation des deux populations le long des deux premiers axes, représentant à eux deux un peu moins de 12% de la variance totale. Cette première preuve qu'il existait de fortes différences inter-populationnelles dans les profils de méthylation globaux a été renforcée par la détection de 12,050 DMS pour lesquels une différence significative de méthylation pouvait être observée. Un résultat inattendu toutefois a été la réalisation qu'il existait une très forte dissymétrie dans les profils de méthylation des DMS : presque 75% d'entre eux montraient un niveau de méthylation en moyenne plus élevé chez les individus d'origine africaine. Ce résultat est d'autant plus surprenant que l'identification des bases génétiques de ces différences nous ont permis d'identifier une forte composante génétique à cette dissymétrie. En effet, lorsque les DMS étaient sous contrôle génétique, l'allèle responsable d'une plus forte méthylation était très généralement plus fréquent dans la population d'origine africaine que dans la population d'origine européenne, résultant dans une méthylation plus élevée chez les Africains. Ce déséquilibre, statistiquement significatif, est d'autant plus troublant qu'il est associé à des enrichissements en fonctions biologiques totalement différents. Les gènes plus méthylés chez les Européens relativement aux individus Africains, sont enrichis en fonction liées au système immunitaire, tandis que les gènes plus méthylés chez les Africains sont associés à la membrane plasmique et à l'interface de la cellule avec son environnement.

L'étude des bases génétiques de la méthylation que nous venons de mentionner a été conduite localement, en *cis* et à l'échelle du génome, en *trans*. Il me semble pertinent de discuter les résultats obtenus en *trans* : en effet, afin de réduire le nombre de tests, nous nous sommes restreints à associer la méthylation de tous les sites CpGs aux variants génétiques à proximité des facteurs de transcription (TF) activement exprimés dans les monocytes, ainsi qu'aux variants génétiques que nous avons pu détecter comme *cis*-eQTL des-dits TF. L'idée sous-jacente est que les SNPs contrôlant l'expression des TF pourraient être associés à la méthylation des sites CpG à proximité des sites de fixation de ces mêmes TF, puisqu'il a été démontré que l'activité des TF peut modifier la méthylation localement. Nos résultats nous ont permis de confirmer la véracité de cette idée puisque une écrasante majorité des sites CpG contrôlés génétiquement en *trans* tombaient au sein d'un site de fixation d'un TF. Un résultat intéressant toutefois vient du fait que nous avons identifié un variant, situé à proximité du gène codant CTCF, qui régule le niveau de méthylation de l'ADN de 30 sites CpG. Or bien que tous ces CpG soient localisés au sein d'un site de fixation d'un TF, seul 9 tombent dans un site de fixation spécifique à CTCF, les 21 autres étant localisés dans un site de fixation d'autres TF. Ce résultat nous semble être un argument en faveur d'un modèle décrit par Zaret et Caroll, celui des facteurs pionniers (*pioneer transcription factor*) (Zaret & Carroll, 2011). Ce modèle suggère que certains TF ont une activité pionnière qui leur permet d'accéder à leurs sites de fixation au sein d'états de condensation de la chromatine plus élevés que la moyenne des autres TF. Ainsi, la fixation des TF pionniers permettrait à l'ADN de s'ouvrir et d'être accessible à une seconde vague de TF.

Nous avons ensuite examiné les relations entre variabilité de la méthylation de l'ADN et variabilité de l'activité transcriptionnelle, à la fois à l'état basal (eQTM), et en réponse

à une stimulation du système immunitaire (reQTM). Ceci nous a permis d'établir une carte des eQTM et des reQTM, et d'identifier les régions génomiques enrichies en eQTM ou reQTM, en fonction de leur effet, positif ou négatif, sur l'expression. Par ailleurs, nos résultats ont montré un très fort enrichissement des eQTM en sites CpG sous contrôle génétique (33 fois plus que pour l'ensemble des sites CpG). Cet enrichissement apporte un doute quant au rôle de la méthylation sur l'expression, qui peut être passif ou actif. En effet, la corrélation observée entre méthylation et expression pourrait être la conséquence du contrôle indépendant de ces deux variables par un même et unique variant génétique, auquel cas la méthylation serait alors un marqueur de l'expression sans jouer de rôle causal dans sa régulation (figure 3.1). Afin de résoudre cette situation nous avons utilisé une méthode de médiation qui nous a permis d'identifier les sites CpG dont le niveau de méthylation régulait de manière causale l'expression des gènes, ou leur réponse transcriptionnelle à l'activation immunitaire. Nous avons ainsi identifié qu'environ 20% des gènes associés à un eQTM ou un reQTM étaient sous le contrôle direct de la méthylation de l'ADN. Ce résultat est particulièrement intéressant puisqu'il nous permet d'émettre l'hypothèse que la méthylation de l'ADN à ces sites CpG pourraient être responsable de phénomènes d'adaptation locale. Dans le cas d'un changement extraordinaire et brusque de l'environnement, on peut imaginer que la méthylation de l'ADN soient modifiée rapidement en de tels sites CpG, résultant en des modifications des profils de transcription pour s'adapter au nouvel environnement. Dans le cas d'une persistance de ce nouvel environnement, les modifications de la méthylation de l'ADN pourraient par la suite être fixées par des modifications de la séquence d'ADN. Une idée qui sera discutée plus en avant dans la discussion générale de cette thèse.

En conclusion, notre étude a révélé l'ampleur des différences de méthylation entre des groupes d'individus d'origine ethnique différente, ainsi que les bases génétiques de ces différences. Elle a aussi permis d'affirmer que la méthylation de l'ADN pouvait jouer un rôle causal dans les processus de régulation de l'expression, en particulier lors de la réponse transcriptionnelle à l'infection, dépassant ainsi un simple statut de marqueur de l'expression.

Chapitre 6

Résultat 2 : À la découverte de la spécificité cellulaire et de la causalité des effets de la méthylation de l'ADN dans la régulation de l'activité des gènes de l'immunité dans le sang.

6.1 Contexte

Comme nous l'avons vu dans les chapitres précédents, que ce soit dans l'introduction ou au travers du premier article, un nombre croissant d'études s'accordent à donner un rôle prépondérant à la méthylation de l'ADN dans la régulation de l'expression des gènes, particulièrement dans le contexte du système immunitaire. En effet, le sang étant un des tissus les plus facilement accessibles, c'est le tissu de prédilection des études conduites chez l'espèce humaine. Comme les cellules impliquées dans le système immunitaire circulent majoritairement dans le sang et que l'étude de l'immunité concentre un grand nombre de promesses nous permettant de lutter contre les pathogènes qui nous entourent, on comprend facilement l'attention portée à la méthylation des cellules sanguines. Toutefois, comme discuté au cours des chapitres introductifs il existe une forte spécificité cellulaire des profils d'expression et de méthylation de l'ADN. En effet, puisque toutes les cellules de notre organisme partagent la même séquence d'ADN mais exercent des fonctions pouvant être infiniment différentes, il apparaît clairement qu'au sein de chaque type cellulaire sera mis en place au cours de sa différenciation les profils d'expression et de méthylation spécifiques à sa fonction. Ce double constat complique donc l'étude des relations existant entre méthylation de l'ADN et expression dans les tissus hétérogènes, tels que le sang. En effet, en attendant le développement et (surtout) la réduction du coût des techniques de séquençage de cellule unique, le meilleur moyen d'identifier des eQTM dans un tissu hétérogène repose sur l'utilisation de données de cytométrie en flux qui permettent d'ajuster les modèles linéaires d'association entre méthylation et expression par les proportions en types cellulaires. Néanmoins, ce type d'analyse requiert donc des bases de données combinant des profils d'expression, de méthylation et de cytométrie en flux pour un grand nombre d'échantillons, et sont donc assez rares.

Dans ce second projet, nous avons tiré profit d'une ressource extrêmement complète : le projet *Milieu Intérieur*, qui cherche à comprendre et définir la variabilité de la réponse

du système immunitaire dans la population générale, et à déterminer les facteurs génétiques, épigénétiques et environnementaux contribuant à l'hétérogénéité des phénotypes immunitaires. À cette fin, du sang a été collecté chez une cohorte de 1,000 sujets sains, stratifiés par sexe (500 hommes / 500 femmes) et âge (200 sujets par décennie de 20 à 69 ans), et originaires de France Métropolitaine sur 3 générations. Entres autres, des données de génotypage, de méthylation, de cytométrie en flux et d'expression de 560 gènes de l'immunité ont été produites. C'est à partir de ces données que nous avons posé les bases de cette seconde étude, à savoir étudier le rôle causal de la méthylation dans la régulation de l'expression des gènes de l'immunité dans le sang. L'ensemble des données disponibles nous a en effet permis de construire un graphe de causalité représentant l'étendue des interactions pouvant exister entre la méthylation, l'expression et l'ensemble des variables à notre disposition. De plus, la conception de cette cohorte nous a donné le cadre idéal pour tester l'effet de l'âge et du sexe sur les associations entre méthylation et expression. Enfin, nous avons tiré parti des données de cytométrie en flux pour déterminer l'ampleur de la spécificité cellulaire de la régulation de l'expression par la méthylation.

6.2 Article 2

Understanding the cell-specific and causal effects of DNA methylation in immune gene regulation of whole blood

Lucas T. Husquin[1,2], Jacob Bergstedt[1], Maxime Rotival[1], Etienne Patin[1], Julia L. MacIsaac[3], Michael S. Kobor[3], Lluis Quintana-Murci[1]

[1]Human Evolutionary Genetics Unit, Institut Pasteur, UMR2000, CNRS, Paris 75015, France

[2]Sorbonne Universités, École Doctorale Complexité du Vivant, 75005 Paris, France

[3]Department of Medical Genetics, University of British Columbia, Centre for Molecular Medicine and Therapeutics, BC Children's Hospital Research Institute, Vancouver, Canada

Abstract

DNA methylation variation can result from both environmental and genetic factors, and there is growing evidence that such variation can directly affect gene expression, in particular in the context of the immune system. In this study, we used a comprehensive data resource to build a causal graph and consequently investigate the causative effect of DNA methylation on the expression of 560 immune-related genes. We find that the transcription rate of a vast majority of immune genes (62%) is directly regulated by DNA methylation of at least one nearby CpG site. Furthermore, we observe that DNA methylation is the main variable explaining gene expression variability, relative to a broad range of environmental, intrinsic and genetic factors. Finally, using detailed cytometric data from 69 immune cell types, we examine the cellular specificity of the associations detected, and find that a large proportion of DNA methylation-gene expression associations are cell-type specific. However, we do not find any major age- or sex-specific effect of DNA methylation on gene expression, an observation that is at odds with the known, major effect of age on DNA methylation. Overall, our findings help refining the regulatory regions of the genome and potentially uncover new ones; regardless, our study sheds new light onto the causal, cell-specific role of DNA methylation, and of the promoter-interacting regions, in regulating the transcriptional activity of immune related genes.

Introduction

All information needed to construct an organism is encoded within its DNA molecules. In a sense, organisms can be considered as the organic interface by which the DNA molecules manipulate the world in the pursuit of replication (Dawkins 1976). The main way organisms carry out the instructions within DNA molecules is through the process of transcription, in particular of the coding regions. As such, gene expression is paramount in determining what we are, during the early developmental stages of life (Kornberg and Tabata 1993; LaFlamme 2014), and our capacity to adapt to varying environments (Tung and Gilad 2013; Fave et al. 2018). As a consequence, clear differences in gene expression are observed within a given species: between cells, tissues, individuals or populations (Quach et al. 2016; Nedelec et al. 2016; Chalancon et al. 2012; Raj and van Oudenaarden 2008; Breschi et al. 2016). Yet, despite a large number of studies have investigated the extent of this variation and its drivers over the past decade, the respective contribution of its drivers remains to be explored in further detail. Specifically, studies in whole-blood have delineated a broad effect of age and sex on gene expression (Piasecka et al. 2018; Jansen et al. 2014; Peters et al. 2015), and genetic variation has been long known to be a major driver of gene expression variability (Cheung et al. 2003; Quach et al. 2016; Albert and Kruglyak 2015; Consortium 2015; Võsa et al. 2018). In this context, the mapping of expression quantitative trait loci (eQTLs; genetic variants that affect gene expression variation) has become an important approach to dissect the regulatory mechanisms of gene expression (Montgomery and Dermitzakis 2011). Yet, a large fraction of the variance in gene expression, both across individuals and populations, cannot be attributed to genetic nor to other environmental and intrinsic factors, and remains unexplained (Fairfax et al. 2014; Barreiro et al. 2012; Piasecka et al. 2018).

In this context, DNA methylation may provide an additional layer for variation in gene expression (Bonder et al. 2014; Pai, Pritchard, and Gilad 2015; Husquin et al. 2018; Wagner et al. 2014). Following a canonical model, high levels of methylation at promoter regions are often associated with low gene expression, while methylation of gene bodies tends to be more associated with active expression (Wagner et al. 2014; Yang et al. 2014; Jjingo et al. 2012; Gutierrez-Arcelus et al. 2013). Furthermore, DNA methylation of various regulatory regions, such as enhancers, promoter-interacting regions (PIRs) or Transcription Factor Binding Sites (TFBS), has been proposed as a mode of action through which it could directly affect transcriptional activity (Husquin et al. 2018; Zhu, Wang, and Qian 2016; Hu et al. 2013; Yin et al. 2017; Aran and Hellman 2013; Pellacani et al. 2016; Spurrell, Dickel, and Visel 2016). Given that genetic variants associated with DNA methylation variation (meQTLs) have been shown to be enriched in genome-wide association studies (GWAS) hits (Relton and Davey Smith 2012; Richardson et al. 2017; Hannon et al. 2017), there has been an increasing emphasis in dissecting the causal relationships linking DNA methylation, gene expression, and DNA sequence variation (Pai, Pritchard, and Gilad 2015; Gutierrez-Arcelus et al. 2013; Bonder et al. 2017). Causality analyses - including the likes of mediation, mendelian randomization, or the use of causal graphs - are particularly useful to disentangle the respective contribution of all the factors involved in multifactorial processes, such as the regulation of gene expression (Pearl 2010).

Given the marked tissue specificity of DNA methylation (Pai et al. 2011; Zhou et al. 2017), only 4 to 42% of DNA methylation-gene expression associations have been found to be replicable across tissues (Bonder et al. 2014; Gutierrez-Arcelus et al. 2015). In consequence, recent studies focused on data collected either from virus-transformed lymphoblastoid cell lines, or from specific cell types, such as monocytes. Despite handling the tissue specificity of DNA

methylation, this led to results that could reflect, at least partially, epigenetic changes induced by cell immortalization, or that faded away from a realistic model of the real-life interactions occurring in whole blood (Sugawara et al. 2011; Husquin et al. 2018; Bonder et al. 2017). Conversely, in the context of whole blood, which is a highly heterogenous tissue in terms of cell populations, several deconvolution methods have been developped to assess the cell composition of whole blood based on DNA methylation variation (Houseman et al. 2012; Teschendorff and Relton 2018; Rahmani et al. 2018; Jaffe and Irizarry 2014). However, in studies that have explored DNA methylation from whole blood, varying degrees of importance have been granted to the impact that heterogeneity in blood cell composition has on DNA methylation variation (Lam et al. 2012; Fagny et al. 2015; Bell et al. 2018; Richardson et al. 2017). Thus, the study of the causal relationships between DNA methylation and gene expression in the setting of a heterogeneous tissue, such as whole blood, is still lacking.

As the immune system is the primary interface with the human pathogenic environment, the study of the regulation of the transcription of immune-related genes offers a unique opportunity to explore the interplay between the genome, the epigenome and environmental cues (Piasecka et al. 2018). Here, we built upon the *Milieu Intérieur* data resource, an integrative approach that combines high-dimensional data, including epigenetic, genetic, cytometric, transcriptional and environmental data, obtained from the same individuals (Thomas et al. 2015). We analyzed DNA methylation profiles of more than 850,000 CpG sites for whole blood, a dataset that was combined with data from the proportions of 69 immune cell populations using flow cytometry (Patin et al. 2018). This allowed us to investigate the causal impact of DNA methylation on the regulation of gene expression, while adjusting for intrinsic, genetic and environmental confounders. We also leveraged our flow cytometry data to examine the cellular specificity of the

associations uncovered. Using a system-level approach, we assessed the extent to which inter-individual variation in DNA methylation, in interaction with genetic variation, impact transcriptional activity related to immune activity.

Results

We analyzed DNA methylation variation at > 850,000 CpG sites across the genome, from whole blood originating from 962 healthy European volunteers divided in 5 decades of life (from 20 to 69 years old), and stratified by sex (484 men, 478 women, see Table 1). After normalization and filtering (see "Materials and methods"), we retained a final dataset of 629,175 methylation sites. Furthermore, the genotyping profiles, gene expression data for 560 immune-related genes and the proportions of 69 different cell-types using standardized flow cytometry were available for the same individuals (Additional File 1: Table S1, Fig. 1a) (Patin et al. 2018; Piasecka et al. 2018). Finally, all individuals answered a thorough survey ranging from health-markers to socio-economic factors (Thomas et al. 2015). We used this comprehensive experimental design to define the respective effects of age, sex, environmental, genetic and epigenetic factors, and of inherent variation in immune cell populations on the inter-individual variability of immune-related gene expression.

We hypothesized the causal structure relating variation in DNA methylation to variation in gene expression, summarized in a causal structural equation model (Pearl 2010)(Fig. 1b). It includes the critical components of genetic and cellular heterogeneity, intrinsic factors such as age and sex, as well as the environmental factor of tobacco smoking status, which is known to heavily impact both the DNA methylome and gene expression (Charlesworth et al. 2010; Harris et al. 2017; Zheng et al. 2018; Husquin et al. 2018). In addition, we included cytomegalovirus (CMV) seropositivity, which we have previously detected to also have a marked impact on DNA methylation (Patin et al. 2019). By adjusting for all of these factors, we are able to quantify the strength of the direct link between methylation and gene expression. This estimate is a good

approximation of the causal effect of DNA methylation variation on gene expression variation, given that our causal model is a good approximation of reality. Consequently, the interpretation of all our results only makes sense in the light of the causal links illustrated by this graph.

To adjust for genetic variation and cellular heterogeneity, we first estimated the subsets of cell proportions and local genetic variants that were associated to the expression of each gene. These subsets were estimated, for each gene, by fitting a joint linear model with all cell proportions, or all local variants, and selecting the associated predictors using stability selection (Meinshausen and Bühlmann 2010; Shah and Samworth 2013). This algorithm is an ensemble algorithm that aggregates the selection of important predictors through elastic net regression on each of 100 of subsamples of the data (see "Materials and methods"). The algorithm includes two parameters (q and alpha) that we tuned to maximize our power to detect at least a major cell type, or a major genetic variant per gene (Fig. 2a). This allowed us to detect specific cell types (0 to 9 per gene [median = 4 +/- 2], Additional File 2: Figure S1a) or specific genetic variants (0 to 10 per gene [median = 1 +/- 2], Additional File 2: Figure S1b) whose proportion in whole blood or variability, respectively, best explained gene expression.

Using these variables, we next built the linear regression model of the relationships between gene expression and DNA methylation, for each of the 542 autosomal immune-related genes, adjusting for age, sex, smoking, CMV seropositivity, selected genetic variants and cell types (see "Materials and methods"). We ran this model for each CpG site located within a 100kb of the gene of interest (for a total number of 52,614 tested CpG sites). Associations were considered significant if they passed a *P*-value threshold determined using 100 permutations (FDR = 5%, $P < 2.9 \times 10^{-3}$). In doing so, we detected 334 genes (62% of the total number of genes tested),

whose transcriptional variation was associated which at least one CpG site (referred to as eQTMs), for a total of 2,939 significant associations.

We next aimed to functionally characterize the detected eQTMs, and used the Illumina annotation manifest, along with the Epigenome RoadMap Consortium for whole blood-specific promoter and enhancer regions (Roadmap Epigenomics et al. 2015), the Gene Transcription Regulation Database for Transcription Factor Binding Sites (TFBS) (Yevshin et al. 2019) and cell-specific promoter interacting region (PIRs) detected by promoter capture Hi-C (Javierre et al. 2016). We used this comprehensive annotation of CpG sites to detect a strong enrichment of eQTMs, compared to the global location of all tested CpG sites, in enhancers (OR ~ 2.7, $P < 2.4 \times 10^{-126}$) and, to a lesser extent, in PIRs (OR ~ 1.2, $P < 7.8 \times 10^{-5}$). Because CpG sites whose state of methylation repress gene expression tend to be located in different regions to those that activate expression (Bell et al. 2011), we investigated positive-eQTMs (DNA methylation with activating effects) and negative-eQTMs (DNA methylation with repressing effects), separately. In promoters, we detected a depletion (OR ~ 0.8, $P < 4.4 \times 10^{-3}$) and an enrichment (OR ~ 1.2, $P < 6.0 \times 10^{-5}$) of positive- and negative-eQTMs, respectively, as expected by the canonical model (Fig. 2b). In turn, UTR regions, which have previously been shown to carry eQTMs (Husquin et al. 2018; Grundberg et al. 2013), are enriched in positive-eQTMs exclusively, particularly 5'UTR (OR ~ 1.5, CI = [1.33-1.74], $P < 3.0 \times 10^{-9}$, Fig. 2b).

Besides, to improve the characterization of the genomic location of eQTMs, and, ultimately, better understand the mode of action of regulatory CpG sites, we compared their proximity to TFBS to that of all tested CpG sites. To focus on hotspots of transcription factor activity, we first identified the regions of the genome where 3, 5 or 10 TFBS overlapped (see "Materials and

methods", Additional File 2: Figure S2). This allowed us to detect an enrichment of eQTMs in the vicinity of hotspots of TF activity, enrichment that increases as a function of the number of overlapping TFBS (Fig. 2c). These enrichments were not restricted to the close vicinity of TFBS overlapping regions, as similar results were found for larger windows around such regions (Additional File 2: Figure S3).

Finally, to understand the relative importance of DNA methylation in explaining the variability of gene expression, as compared to the other variables, we estimated the relative effect sizes of all the variables included in the causal model - including selected cell types, genetic variants, sex, age, CMV status and significant methylation probes. Among the 334 genes for which we detected at least one significant eQTM, we found that DNA methylation and variability in immune cell proportions were the two variables explaining most of the variance of gene expression in 148 (44.3% of tested genes, mean variance explained = 17.7% +/- 9.2) and 124 (37.1%, mean = 16.7% +/- 10.1) genes, respectively (Fig. 2d). Furthermore, looking at the overall amount of variance explained, summing across all genes with at least one eQTM, we confirmed the predominant importance of DNA methylation and cellular proportions in explaining gene expression (Additional File 2: Figure S4). Interestingly, although genetics was the most prominent variable for only 32 genes (9.6%, mean = 22.5% +/- 15.0), for these cases effect sizes tend to be very high. For example, more than 50% of the variance in expression of *LILRA3*, *IFIT2* and *HLA.C* was found to be explained nearby genetic variants (59.7%, 57.8% and 53.9%, respectively), as previously observed (Piasecka et al. 2018).

Nevertheless, a large portion of the variance of gene expression remained unexplained, ranging from 33.1% to 97.7%, highlighting the progress that still can be achieve to better understand the

underlying causes of gene expression. Focusing on the 176 genes for which we did not find any significant eQTM, we found that cellular proportions explained most of the variance of expression for 81 genes (46.0%), far more than genetic variants and sex (18.2% and 16.5%, respectively)(Additional File 2: Figure S5-S6). We also observed that the position of the CpG sites, relatively to the promoters, enhancers and PIRS, appeared to be linked with its distance to a TFBS overlap region, indicating the coordinated role of these regulatory regions (Fig. 2e). However, the quantitative impact of DNA methylation on gene expression did not seem to be affected by its distance to the different regulatory elements of the genome that we tested (Fig.2e, Additional File 2: Figure S7).

The processes that link methylation to gene expression in whole blood take part within immune cells. In that sense the magnitude of the effects of eQTMs is the average of the magnitude of the effects specific to each cell-type. Fortunately, the availability of the 69 cell-type proportions made it possible to directly investigate the cell-specific dependency of gene expression on methylation. For each gene, we included an interaction term between methylation and each cell-type previously selected by the stability selection algorithm (one model per interaction term, see "Materials and methods"). Strikingly, following this procedure, we found 1,217 associations (representing 41.4% of all significant eQTMs) that showed a cell-specific effect with at least one cell-type. As a general example, we show the effect of the methylation at cg12271317, located on chromosome 4, on the levels of expression of *LEF1*, which depend on the amount of CD8+ naive cytotoxic T cells (Fig. 3a). As for eQTMs, we characterized the genomic location of the CpGs sites whose DNA methylation show cell-specific effects on gene expression (referred to as cell-eQTMs). With respect to all CpG sites within 100 kb of immune genes, cell-eQTMs were highly enriched in enhancers (OR ~ 3.6, $P < 4.1 \times 10^{-98}$), promoters (OR ~ 1.8, $P < 5.2 \times 10^{-21}$), PIRs

(OR ~ 10.0, $P < 9.5 \times 10^{-266}$) and TFBS overlapping regions (Additional File 2: Figure S8-S9). Nevertheless, to go further and get insights about the mechanisms underlying cell-specific regulation of gene expression, we compared cell-eQTMs to all detected eQTMs. While we did not find significant enrichments in enhancers, promoters or TFBS overlapping regions, we observed that cell-eQTMs were preferentially located in enhancers (OR ~ 1.7, $P < 1.7 \times 10^{-11}$), and especially in PIRs (OR ~ 4.8, $P < 8.2 \times 10^{-70}$), highlighting the peculiar importance of such promoter interacting regions in regulating gene expression in a cell-specific context, relatively to other regulatory regions (Fig.3b,c).

Specifically, we found that a few major cell types, such as CD4 T-cells, had a paramount influence on the way DNA methylation regulates the expression of several genes such as *LCK*, *CD3E* or *CD3D* (Fig.3d). Interestingly, these genes were mostly involved in the specific processes of T-cells activity: *LCK* encodes for a protein that plays a central role in the selection and maturation of developing T-cells (Ito et al. 2003; Vogel and Fujita 1995), and both *CD3E* and *CD3D* encode subunits of the T-cell receptor-CD3 complex, which plays an important role in coupling antigen recognition to several intracellular signal-transduction pathways (de Saint Basile et al. 2004). Furthermore, focusing on the most significant results of our interaction analysis, two examples stand out: in B-cells methylation at cg27565966 regulates the expression of *CD19*, a gene known to encode a co-receptor for the B-cell antigen receptor complex (BCR) on B-lymphocytes (Interacting-Pvalue = 5.4×10^{-13}) (Tedder, Zhou, and Engel 1994). On the other hand, in T-cells methylation at cg02625086 regulates the expression of *LEF1*, whose end product is a known to bind to a functionally important site in the T-cell receptor-alpha enhancer, thereby conferring maximal enhancer activity (Interacting-Pvalue = 4.3×10^{-12})(Additional File 4: Table S2) (Willinger et al. 2006). These examples emphasize the importance of such

interaction analyses to characterize relevant effect of DNA methylation on gene expression in a cell-specific context.

Finally, we tested the interaction with the other variables initially included in the general model for mapping eQTMs, such as sex, age, CMV status and genetic variants. However, we found little to no significant interactions between DNA methylation and these variables (Additional File 2: Figure S10). We found only 51 significant sex-eQTMs and 19 snp-eQTMs, figures that were unfortunately too low to conduct any enrichment analyses as we did for cell-eQTMs, but that provide once again a good proof of concept of this approach.

Discussion

This study sheds new light onto the impact that DNA methylation variation has on the regulation of gene expression, in the context of immune-related genes. Population level investigations of biomolecular relationships and how they depend on the tissue and intrinsic and environmental factors necessarily require observational data. The comprehensive data collected for the Milieu Interieur resource made it possible to build models that capture this context. Such models reflect our underlying assumption of how the various factors in the system are dependent on each other, specified in a causal graph (Fig 1b). In the context of this model, we estimate the causal effect of DNA methylation on gene expression and identify the CpG sites directly affecting gene expression, thus allowing a better characterization of the regions of the genome involved in the regulatory processes of transcriptional activity (Bonder et al. 2017; Husquin et al. 2018; Pai, Pritchard, and Gilad 2015). We believe that the causal structural equation model framework is well suited for genetic epidemiology studies such as eQTM mapping – but also eQTLs, meQTLs and, more generally, all observational studies of molecular phenotypes.

Furthermore, the analysis of the contribution of each variable of the model to the overall gene expression patterns allowed us to quantify the extent to which DNA methylation regulates gene expression, while adjusting for cellular heterogeneity, age, sex and CMV serostatus. From a quantitative perspective, we found that DNA methylation and immune cellular heterogeneity in whole blood are the main variable regulating gene expression. Although genetic factors appear to explain a high fraction of gene expression variance for specific genes or pathways, we found an overall moderate influence of genetics in the regulation of the overall expression of immune

genes in whole blood, as previously reported (Brodin et al. 2015; Orru et al. 2013; Carr et al. 2016; Roederer et al. 2015).

Our study also represents a comprehensive investigation of the role of DNA methylation in regulating the transcriptional activity of immune-related genes in the context of whole blood. This systems immunology approach allowed us to quantify the link between DNA methylation and gene expression in a cell specific context, while simultaneously integrating intrinsic, environmental and genetic factors known to play a role in the regulation of gene expression (Glass et al. 2013; Peters et al. 2015; Piasecka et al. 2018; Gershoni and Pietrokovski 2017; Grath and Parsch 2016). Furthermore, our age- and sex-stratified cohort allowed us to explore if interactions between these intrinsic variables and the numerous epigenetic factors identified affect immune genes transcription. Surprisingly, we did not find any significant AGE × CpG interaction, contrary to some reports suggesting an effect of age on the hypermethylation of CpG islands that surround tumor repressor genes, or other inflammatory-related genes (Marttila et al. 2015; Bell et al. 2012; Jones and Baylin 2002; Issa 2000). This result, which might be due to our focus on immune functions, suggests that the effects of DNA methylation on the expression of immune gene, at least, are constant across age groups.

Finally, we found that almost half of the detected associations between DNA methylation and gene expression showed CELL × CpG interactions. This indicates a strong cellular variability in the epigenetic control of gene expression, a result that is consistent with the strong tissue specificity of DNA methylation itself (Pai et al. 2011). The fact that we find examples of cell-specific regulation of gene expression for various genes relevant with the studied cell types brings striking evidence for the pertinence of such interaction analyses – especially when

combined with the embedding of our analyses in a causal structural equation model framework – that, in the end, allow for a more precise capture of genuine DNA methylation-gene expression associations.

Materials and Methods

The Milieu Intérieur Cohort. The Milieu Intérieur Project includes 1,000 healthy donors (500 men and 500 women) aged 20–69 y old equally distributed across five decades of life (200 individuals per decade). A more thorough description of the cohort can be found in (Thomas et al. 2015).

DNA Genotyping and Imputation. All individuals were genotyped at 719,665 SNPs on a HumanOmniExpress-24 BeadChip (Illumina) (Piasecka et al. 2018). The coverage of rare functional variants was increased by genotyping an additional 245,766 SNPs on a HumanExome-12 BeadChip (Illumina). After strict quality control filters and merging the two datasets, a total of 723,341 SNPs was retained. Finally, genotype imputation with IMPUTE v.2 (Howie, Donnelly, and Marchini 2009) was performed, using the 1000G Project imputation reference panel (Genomes Project et al. 2010). These steps led to a final dataset of 5,265,361 SNPs that were used for all subsequent analyses.

DNA methylation profiling and data normalization. Genomic DNA was extracted from whole blood using a phenol/chloroform protocol followed by ethanol precipitation. The DNA was then bisulfite converted, and BC-DNA was then processed using the Illumina Infinium MethylationEPIC BeadChip Kit (Illumina, San Diego, CA) to obtain the methylation profile of each individual at more than 850,000 CpG sites genome-wide. In total, 1,088 samples were hybridized with the EPIC array, of which 108 were discarded because of bad quality, leaving 962 unique samples and 18 technical replicates. We removed any technically unreliable probes: (i) potentially cross-hybridizing probes (83,635 probes), (ii) those located on the X and Y

chromosomes (19,681 probes), and (iii) probes overlapping SNPs that present a frequency higher than 1% in at least one of the studied population (135,820 probes). These SNPs were chosen based on our own genotyping dataset, as well as on the 1000 Genomes project (Genomes Project et al. 2010). To control for the quality of the probes and samples, we filtered out individuals with > 5% of probes associated with a detection P value > 10^{-3}, and then, probes with a detection P value > 10^{-3} in one or more individuals (27,551 probes). Following this filtering process, 629,175 of the original 866,836 sites on the array were retained, for 484 men and 478 women.

We calculated methylation levels from raw data, using the R Bioconductor lumi package (Du, Kibbe, and Lin 2008). Given that the M value has been shown to provide better detection sensitivity than β values at extreme levels of modification (Du et al. 2010), we used the M value to run all statistical analysis unless otherwise stated. Note that in some instances of the text and figures, β values are reported for ease of clarity and interpretation. M values were then adjusted for background noise with the normal-exponential using out-of-band probes (noob) from the R Bioconductor minfi package (Aryee et al. 2014). Next, normalization for color bias was performed using lumiMethyC with the "quantile" method, and for methylated/unmethylated intensity variation using the lumiMethyN with the "ssn" method 2008(Du, Kibbe, and Lin 2008). Finally, we corrected for technical differences between type I and type II assay designs, by performing peak-based correction (Dedeurwaerder et al. 2011). To correct for known batch effects and potential hidden confounders, we used the sva function from the sva Bioconductor package (Leek et al. 2012) with age, sex, smoking and CMV infection as variables of interest.

Expression quantitative trait methylation (eQTM) analysis. To identify associations between DNA methylation levels and gene expression of nearby genes, we leveraged expression data

obtained from the same individuals, as previously described (Urrutia et al. 2016). Briefly, a specific chloroform-free one-step protocol based on a modified version of the NucleoSpin 96 RNA tissue kit protocol (Macherey-Nagel) was adapted for use with the Freedom EVO integrated vacuum system. RNA concentration was estimated with the Qubit RNA HS Assay Kit (Life Technologies), and RNA integrity was assessed with the Standard RNA Reagent Kit on a LabChip GX (Perkin-Elmer). The RNA Quality Score (RQS) was calculated with LabChip System software, and all samples with an RQS > 4 were processed for gene expression analysis. The NanoString nCounter system, a hybridization-based multiplex assay, was used for the digital counting of transcripts. From each sample, 100 ng of total RNA was hybridized according to the manufacturer's instructions with the Human Immunology v2 Gene Expression CodeSet, which contains 594 endogenous gene probes, 8 negative control probes (NEG A to NEG H), and 6 positive control probes (POS A to POS F) designed against six in vitro- transcribed RNA targets premixed with the CodeSet at a range of concentrations (from 128 to 0.125 fM).

Using such data, we mapped eQTMs (i.e., CpGs whose variation is associated with gene expression) in a window of 100 kb around each gene, including as covariates in the model sex, age, smoking and CMV along with the specific snps and cell types previously selected for each gene:

$$Gene_i \sim CpG_j + \text{Age} + \text{Sex} + \text{Smoking} + \text{CMV} + \text{Snps}_i + \text{CellTypes}_i \qquad (1)$$

where $Gene_i$ is the expression of the *i*th gene, CpG_j the methylation value of the *j*th CpG site in the vicinity of $Gene_i$, and $Snps_i$ and $CellTypes_i$ are the genotypes of the Snps and the proportions of the cell types specifically selected for the *i*th gene.

We applied the Sidak correction to correct for the variability in the number of local CpG sites around genes:

$$Pvalue_{Corrected} = 1 - (1 - Pvalue)^{n_{CpG}}$$

Additionally, to correct for false positive, we computed the mapping of eQTMs on 100 datasets with the M values permuted, and kept, after each permutation, the most significant corrected *Pvalue* per gene (gene-level FDR). We selected the *Pvalue* threshold that provided a 5% FDR ($P = 5.25 \times 10^{-2}$).

Interaction analyses. To investigate the specificity of the detected eQTMs associations, we leveraged the different variables that we had previously included in the eQTM model. More specifically, we repeated the same linear model (1) including an interaction term between DNA methylation and each of the other variables (one model per interaction term). Here is a general example of a model testing for interaction between DNA methylation and age:

$$Gene_i \sim CpG_j * \text{Age} + \text{Sex} + \text{Smoking} + \text{CMV} + \text{Snps}_i + \text{CellTypes}_i \qquad (2)$$

where $Gene_i$ is the expression of the *i*th gene, CpG_j the methylation value of the *j*th CpG site in the vicinity of $Gene_i$, and $Snps_i$ and $CellTypes_i$ are the genotypes of the Snps and the proportions of the cell types specifically selected for the *i*th gene.
For each variable independently, we then extracted the *Pvalues* of the interaction term and corrected them with the Benjamini and Hochberg method to obtain a 5% FDR.

Detection of hotspots of transcription. To identify regions of high importance in the regulation of gene expression, we aimed at identifying regions where TFBS overlapped. Therefore, we used the Gene Transcription Regulation Database (Yevshin et al. 2019) to extract the position of all TFBS that were detected in Homo sapiens. First, we converted the coordinates, provided in build GRCh38, to build GRCh37, corresponding to the coordinates of Illumina annotations for CpG sites on the EPIC array. Then, we used the IRanges R package (REF) to create a coverage of the TFBS with the function *coverage*, and then a sliced object of this coverage with the function *slice* for 3 different values of required overlap: 3, 5 and 10 base pairs, which in return gave us the start and end positions of all regions of the human genome where 3, 5 and 10 TFBS overlapped.

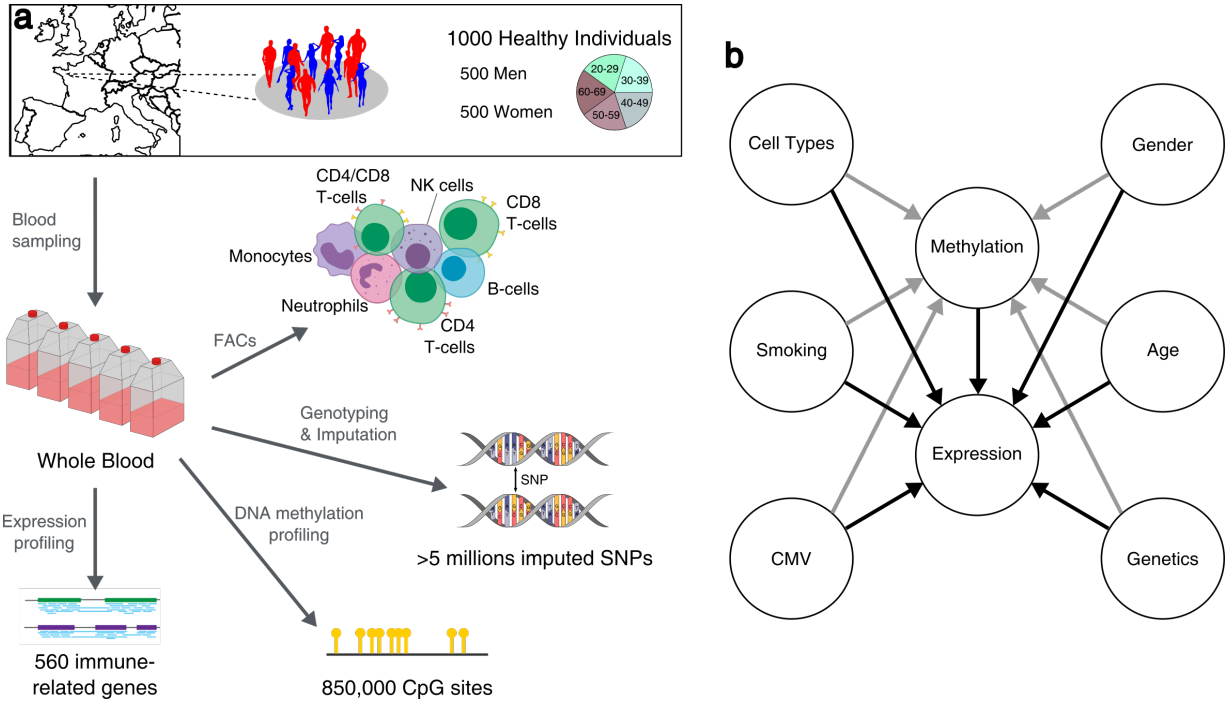

Figure 1. Experimental and statistical design
a, Overview of the Experimental Setting, the transcriptional activity, genotyping and DNA methylation profiling of whole blood from 1,000 healthy individuals equally divided by sex in 5 decades of life was dissected to pinpoint the regulatory processes occurring in the context of immune genes. **b,** Causal structural equation model explaining the variability of gene expression. Genetic polymorphisms, cellular heterogeneity, DNA methylation variation, intrinsic factors (age and sex) along with smoking and CMV seropositivity are included in the causal graph. Dark arrows represent links that will be estimated by the statistical model, grey arrows represent known associations that will not be estimated by our statistical model.

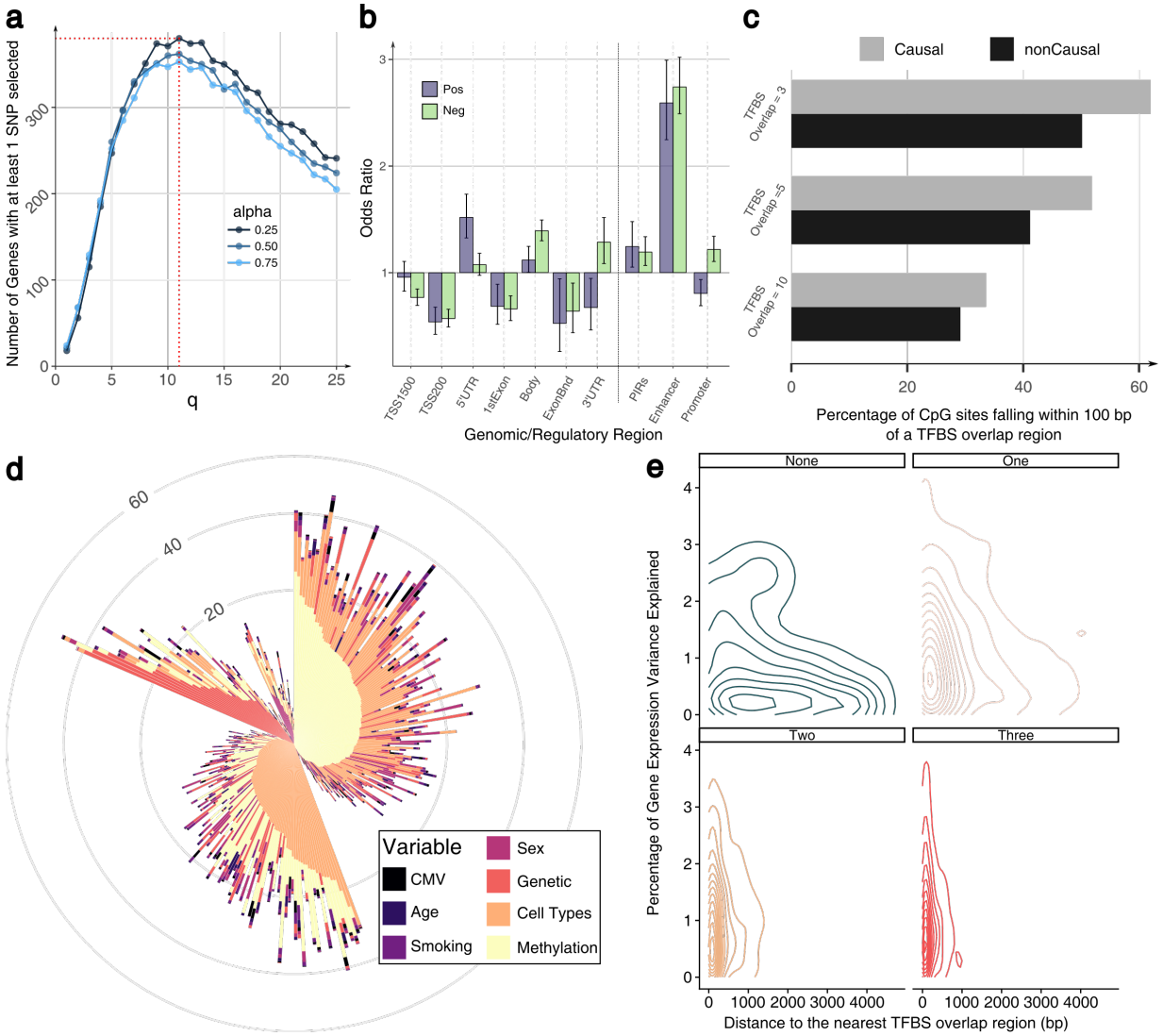

Figure 2. Characterization of DNA methylation effects on gene expression
a, Tuning of q and alpha parameters used in the stability selection model. The y-axis represents the number of genes for which we could detect at least one associated genetic variant, as a function of the q parameter varying from 1 to 25. The three different shades of blue indicate the value of the alpha parameter. The dotted red line indicates the maximum value reached on the y-axis, and the corresponding value of q (q=11, alpha=0.25). **b,** Genomic location of eQTMs, for positive-eQTMs in light purple and negative-eQTMs in light green. Odds ratio and 95% confidence intervals are displayed for positive- and negative-eQTMs, comparing their localization in 7 genomic locations provided by Illumina (TSS1500, TSS200, 5'UTR, 1stExon, Body, ExonBnd and 3'UTR), and in enhancer, promoter and PIRs. Odds ratio were computed against the distribution of the 52,614 tested CpGs. **c,** Enrichment of eQTMs in TFBS overlap regions. The proportions of eQTMs falling within 100 bp of the center of a region where at least 3, 5 or 10 TFBS overlapped is displayed in grey bars. The black bars represent the proportions of the CpGs that were not detected as eQTMs among the 52,614 test CpGs. **d,** Circular bar plot of the percentages of variance explained by the different variables included in the model for each

gene significantly associated with at least 1 eQTM. Each line of the plot represent the total percentage of variance explained for a given gene, with the different colors indicating the respective impact of each variable. Genes are ordered according to the main variable explaining the most amount of variance of their expression. **e,** Contour density plots displaying the percentage of variance explained by the different eQTMs, according to their distance to the nearest TFBS overlap region of at least 10 TFBS. eQTMs were separated in 4 categories, according to whether they fell in 0, 1, 2 or 3 distinct regulatory regions (among PIRs, enhancer and promoter regions).

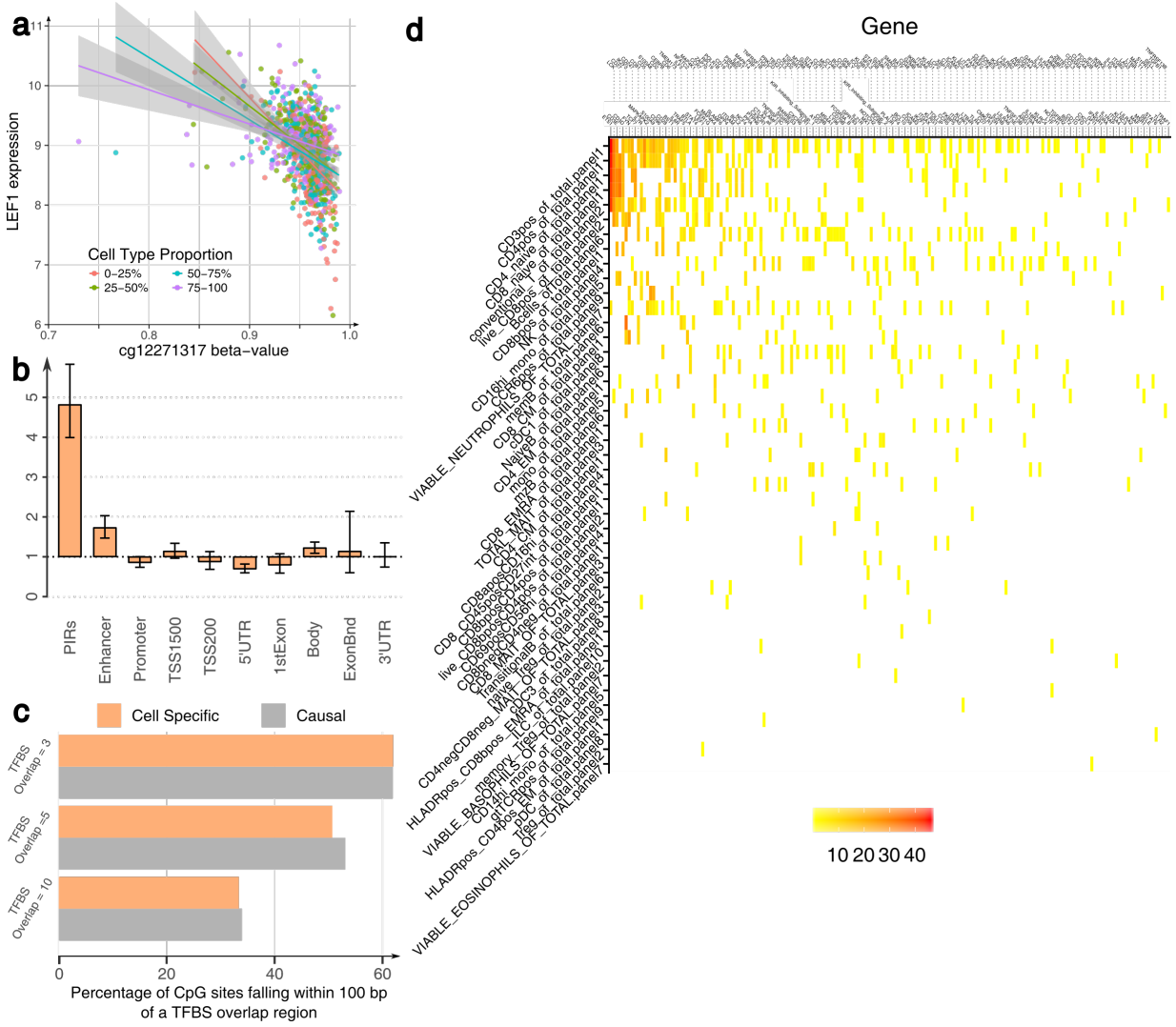

Figure 3. Cell-specific effects of DNA methylation on gene expression
a, Example of a cell-eQTM detected in this study. Lines indicate the fitted linear regression model and grey shades the 95% confidence intervals of these models. The distribution of the expression values of *LEF1* is plotted as a function of the β values of cg12271317, in red, green, blue and purple for individuals with levels of CD8 naïve cells between 0-25%, 25-50%, 50-75% and 75-100%, respectively. **b,** Genomic location of cell-eQTMs, in light orange. Odds ratio and 95% confidence intervals are displayed for the 1,144 cell-eQTMs detected in this study, comparing their localization in 7 genomic locations provided by Illumina (TSS1500, TSS200, 5'UTR, 1stExon, Body, ExonBnd and 3'UTR), and in enhancer, promoter and PIRs. Odds ratio were computed against the distribution of the 1,406 eQTMs for which no significant cell-specific effect was detected. **c,** Enrichment of cell-eQTMs in TFBS overlap regions. The proportions of cell-eQTMs falling within 100 bp of the center of a region where at least 3, 5 or 10 TFBS overlapped is displayed in light orange bars. The black bars represent the proportions of the 1,406 eQTMs for which no significant cell-specific effect was detected. **d,** HeatMap of the number of significant cell-eQTMs, per gene and per cell-type. Shades of yellow and red indicate the number

of eQTMs having a significant interaction with a given cell-type, in rows, for the expression of a given gene, in columns. Blank cells indicate that no significant cell-eQTMs were found for a given gene and a given cell-type.

References

Albert, F. W., and L. Kruglyak. 2015. 'The role of regulatory variation in complex traits and disease', *Nat Rev Genet*, 16: 197-212.

Aran, D., and A. Hellman. 2013. 'DNA methylation of transcriptional enhancers and cancer predisposition', *Cell*, 154: 11-3.

Aryee, M. J., A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry. 2014. 'Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays', *Bioinformatics*, 30: 1363-9.

Barreiro, L. B., L. Tailleux, A. A. Pai, B. Gicquel, J. C. Marioni, and Y. Gilad. 2012. 'Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection', *Proc Natl Acad Sci U S A*, 109: 1204-9.

Bell, C. G., F. Gao, W. Yuan, L. Roos, R. J. Acton, Y. Xia, J. Bell, K. Ward, M. Mangino, P. G. Hysi, J. Wang, and T. D. Spector. 2018. 'Obligatory and facilitative allelic variation in the DNA methylome within common disease-associated loci', *Nat Commun*, 9: 8.

Bell, J. T., A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, and J. K. Pritchard. 2011. 'DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines', *Genome Biol*, 12: R10.

Bell, J. T., P. C. Tsai, T. P. Yang, R. Pidsley, J. Nisbet, D. Glass, M. Mangino, G. Zhai, F. Zhang, A. Valdes, S. Y. Shin, E. L. Dempster, R. M. Murray, E. Grundberg, A. K. Hedman, A. Nica, K. S. Small, Ther Consortium Mu, E. T. Dermitzakis, M. I. McCarthy, J. Mill, T. D. Spector, and P. Deloukas. 2012. 'Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population', *PLoS Genet*, 8: e1002629.

Bonder, M. J., S. Kasela, M. Kals, R. Tamm, K. Lokk, I. Barragan, W. A. Buurman, P. Deelen, J. W. Greve, M. Ivanov, S. S. Rensen, J. V. van Vliet-Ostaptchouk, M. G. Wolfs, J. Fu, M. H. Hofker, C. Wijmenga, A. Zhernakova, M. Ingelman-Sundberg, L. Franke, and L. Milani. 2014. 'Genetic and epigenetic regulation of gene expression in fetal and adult human livers', *BMC Genomics*, 15: 860.

Bonder, M. J., R. Luijk, D. V. Zhernakova, M. Moed, P. Deelen, M. Vermaat, M. van Iterson, F. van Dijk, M. van Galen, J. Bot, R. C. Slieker, P. M. Jhamai, M. Verbiest, H. E. Suchiman, M. Verkerk, R. van der Breggen, J. van Rooij, N. Lakenberg, W. Arindrarto, S. M. Kielbasa, I. Jonkers, P. van 't Hof, I. Nooren, M. Beekman, J. Deelen, D. van Heemst, A. Zhernakova, E. F. Tigchelaar, M. A. Swertz, A. Hofman, A. G. Uitterlinden, R. Pool, J. van Dongen, J. J. Hottenga, C. D. Stehouwer, C. J. van der Kallen, C. G. Schalkwijk, L. H. van den Berg, E. W. van Zwet, H. Mei, Y. Li, M. Lemire, T. J. Hudson, Bios Consortium, P. E. Slagboom, C. Wijmenga, J. H. Veldink, M. M. van Greevenbroek, C. M. van Duijn, D. I. Boomsma, A. Isaacs, R. Jansen, J. B. van Meurs, P. A. t Hoen, L. Franke, and B. T. Heijmans. 2017. 'Disease variants alter transcription factor levels and methylation of their binding sites', *Nat Genet*, 49: 131-38.

Breschi, Alessandra, Sarah Djebali, Jesse Gillis, Dmitri D. Pervouchine, Alex Dobin, Carrie A. Davis, Thomas R. Gingeras, and Roderic Guigó. 2016. 'Gene-specific patterns of expression variation across organs and species', *Genome Biology*, 17: 151.

Brodin, P., V. Jojic, T. Gao, S. Bhattacharya, C. J. Angel, D. Furman, S. Shen-Orr, C. L. Dekker, G. E. Swan, A. J. Butte, H. T. Maecker, and M. M. Davis. 2015. 'Variation in the human immune system is largely driven by non-heritable influences', *Cell*, 160: 37-47.

Carr, E. J., J. Dooley, J. E. Garcia-Perez, V. Lagou, J. C. Lee, C. Wouters, I. Meyts, A. Goris, G. Boeckxstaens, M. A. Linterman, and A. Liston. 2016. 'The cellular composition of the human immune system is shaped by age and cohabitation', *Nat Immunol*, 17: 461-68.

Chalancon, G., C. N. Ravarani, S. Balaji, A. Martinez-Arias, L. Aravind, R. Jothi, and M. M. Babu. 2012. 'Interplay between gene expression noise and regulatory network architecture', *Trends Genet*, 28: 221-32.

Charlesworth, J. C., J. E. Curran, M. P. Johnson, H. H. Goring, T. D. Dyer, V. P. Diego, J. W. Kent, Jr., M. C. Mahaney, L. Almasy, J. W. MacCluer, E. K. Moses, and J. Blangero. 2010. 'Transcriptomic epidemiology of smoking: the effect of smoking on gene expression in lymphocytes', *BMC Med Genomics*, 3: 29.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen, M. Morley, and R. S. Spielman. 2003. 'Natural variation in human gene expression assessed in lymphoblastoid cells', *Nat Genet*, 33: 422-5.

Consortium, G. TEx. 2015. 'Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans', *Science*, 348: 648-60.

Dawkins, Richard. 1976. *The selfish gene* (Oxford University Press: Oxford).

de Saint Basile, G., F. Geissmann, E. Flori, B. Uring-Lambert, C. Soudais, M. Cavazzana-Calvo, A. Durandy, N. Jabado, A. Fischer, and F. Le Deist. 2004. 'Severe combined immunodeficiency caused by deficiency in either the delta or the epsilon subunit of CD3', *J Clin Invest*, 114: 1512-7.

Dedeurwaerder, S., M. Defrance, E. Calonne, H. Denis, C. Sotiriou, and F. Fuks. 2011. 'Evaluation of the Infinium Methylation 450K technology', *Epigenomics*, 3: 771-84.

Du, P., W. A. Kibbe, and S. M. Lin. 2008. 'lumi: a pipeline for processing Illumina microarray', *Bioinformatics*, 24: 1547-8.

Du, P., X. Zhang, C. C. Huang, N. Jafari, W. A. Kibbe, L. Hou, and S. M. Lin. 2010. 'Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis', *BMC Bioinformatics*, 11: 587.

Fagny, Maud, Etienne Patin, Julia L. MacIsaac, Maxime Rotival, Timothée Flutre, Meaghan J. Jones, Katherine J. Siddle, Hélène Quach, Christine Harmant, Lisa M. McEwen, Alain Froment, Evelyne Heyer, Antoine Gessain, Edouard Betsem, Patrick Mouguiama-Daouda, Jean-Marie Hombert, George H. Perry, Luis B. Barreiro, Michael S. Kobor, and Lluis Quintana-Murci. 2015. 'The epigenomic landscape of African rainforest hunter-gatherers and farmers', *Nature Communications*, 6: 10047.

Fairfax, B. P., P. Humburg, S. Makino, V. Naranbhai, D. Wong, E. Lau, L. Jostins, K. Plant, R. Andrews, C. McGee, and J. C. Knight. 2014. 'Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression', *Science*, 343: 1246949.

Fave, M. J., F. C. Lamaze, D. Soave, A. Hodgkinson, H. Gauvin, V. Bruat, J. C. Grenier, E. Gbeha, K. Skead, A. Smargiassi, M. Johnson, Y. Idaghdour, and P. Awadalla. 2018. 'Gene-by-environment interactions in urban populations modulate risk phenotypes', *Nat Commun*, 9: 827.

Genomes Project, Consortium, G. R. Abecasis, D. Altshuler, A. Auton, L. D. Brooks, R. M. Durbin, R. A. Gibbs, M. E. Hurles, and G. A. McVean. 2010. 'A map of human genome variation from population-scale sequencing', *Nature*, 467: 1061-73.

Gershoni, M., and S. Pietrokovski. 2017. 'The landscape of sex-differential transcriptome and its consequent selection in human adults', *BMC Biol*, 15: 7.

Glass, D., A. Vinuela, M. N. Davies, A. Ramasamy, L. Parts, D. Knowles, A. A. Brown, A. K. Hedman, K. S. Small, A. Buil, E. Grundberg, A. C. Nica, P. Di Meglio, F. O. Nestle, M. Ryten, U. K. Brain Expression consortium, Ther consortium Mu, R. Durbin, M. I. McCarthy, P. Deloukas, E. T. Dermitzakis, M. E. Weale, V. Bataille, and T. D. Spector. 2013. 'Gene expression changes with age in skin, adipose tissue, blood and brain', *Genome Biol*, 14: R75.

Grath, Sonja, and John Parsch. 2016. 'Sex-Biased Gene Expression', *Annual Review of Genetics*, 50: 29-44.

Grundberg, E., E. Meduri, J. K. Sandling, A. K. Hedman, S. Keildson, A. Buil, S. Busche, W. Yuan, J. Nisbet, M. Sekowska, A. Wilk, A. Barrett, K. S. Small, B. Ge, M. Caron, S. Y. Shin, Consortium Multiple Tissue Human Expression Resource, M. Lathrop, E. T. Dermitzakis, M. I. McCarthy, T. D. Spector, J. T. Bell, and P. Deloukas. 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: 876-90.

Gutierrez-Arcelus, M., T. Lappalainen, S. B. Montgomery, A. Buil, H. Ongen, A. Yurovsky, J. Bryois, T. Giger, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, I. Padioleau, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. 2013. 'Passive and active DNA methylation and the interplay with genetic variation in gene regulation', *Elife*, 2: e00523.

Gutierrez-Arcelus, M., H. Ongen, T. Lappalainen, S. B. Montgomery, A. Buil, A. Yurovsky, J. Bryois, I. Padioleau, L. Romano, A. Planchon, E. Falconnet, D. Bielser, M. Gagnebin, T. Giger, C. Borel, A. Letourneau, P. Makrythanasis, M. Guipponi, C. Gehrig, S. E. Antonarakis, and E. T. Dermitzakis. 2015. 'Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing', *PLoS Genet*, 11: e1004958.

Hannon, E., M. Weedon, N. Bray, M. O'Donovan, and J. Mill. 2017. 'Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci', *Am J Hum Genet*, 100: 954-59.

Harris, S. E., V. Riggio, L. Evenden, T. Gilchrist, S. McCafferty, L. Murphy, N. Wrobel, A. M. Taylor, J. Corley, A. Pattie, S. R. Cox, C. Martin-Ruiz, J. Prendergast, J. M. Starr, R. E. Marioni, and I. J. Deary. 2017. 'Age-related gene expression changes, and transcriptome wide association study of physical and cognitive aging traits, in the Lothian Birth Cohort 1936', *Aging (Albany NY)*, 9: 2489-503.

Houseman, E. A., W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey. 2012. 'DNA methylation arrays as surrogate measures of cell mixture distribution', *BMC Bioinformatics*, 13: 86.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. 'A flexible and accurate genotype imputation method for the next generation of genome-wide association studies', *PLoS Genet*, 5: e1000529.

Hu, S., J. Wan, Y. Su, Q. Song, Y. Zeng, H. N. Nguyen, J. Shin, E. Cox, H. S. Rho, C. Woodard, S. Xia, S. Liu, H. Lyu, G. L. Ming, H. Wade, H. Song, J. Qian, and H. Zhu. 2013. 'DNA

methylation presents distinct binding sites for human transcription factors', *Elife*, 2: e00726.
Husquin, L. T., M. Rotival, M. Fagny, H. Quach, N. Zidane, L. M. McEwen, J. L. MacIsaac, M. S. Kobor, H. Aschard, E. Patin, and L. Quintana-Murci. 2018. 'Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation', *Genome Biol*, 19: 222.
Issa, J. P. 2000. 'CpG-island methylation in aging and cancer', *Curr Top Microbiol Immunol*, 249: 101-18.
Ito, T., H. Okazawa, K. Maruyama, K. Tomizawa, S. Motegi, H. Ohnishi, H. Kuwano, A. Kosugi, and T. Matozaki. 2003. 'Interaction of SAP-1, a transmembrane-type protein-tyrosine phosphatase, with the tyrosine kinase Lck. Roles in regulation of T cell function', *J Biol Chem*, 278: 34854-63.
Jaffe, A. E., and R. A. Irizarry. 2014. 'Accounting for cellular heterogeneity is critical in epigenome-wide association studies', *Genome Biol*, 15: R31.
Jansen, R., S. Batista, A. I. Brooks, J. A. Tischfield, G. Willemsen, G. van Grootheest, J. J. Hottenga, Y. Milaneschi, H. Mbarek, V. Madar, W. Peyrot, J. M. Vink, C. L. Verweij, E. J. de Geus, J. H. Smit, F. A. Wright, P. F. Sullivan, D. I. Boomsma, and B. W. Penninx. 2014. 'Sex differences in the human peripheral blood transcriptome', *BMC Genomics*, 15: 33.
Javierre, B. M., O. S. Burren, S. P. Wilder, R. Kreuzhuber, S. M. Hill, S. Sewitz, J. Cairns, S. W. Wingett, C. Varnai, M. J. Thiecke, F. Burden, S. Farrow, A. J. Cutler, K. Rehnstrom, K. Downes, L. Grassi, M. Kostadima, P. Freire-Pritchett, F. Wang, Blueprint Consortium, H. G. Stunnenberg, J. A. Todd, D. R. Zerbino, O. Stegle, W. H. Ouwehand, M. Frontini, C. Wallace, M. Spivakov, and P. Fraser. 2016. 'Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters', *Cell*, 167: 1369-84 e19.
Jjingo, D., A. B. Conley, S. V. Yi, V. V. Lunyak, and I. K. Jordan. 2012. 'On the presence and role of human gene-body DNA methylation', *Oncotarget*, 3: 462-74.
Jones, P. A., and S. B. Baylin. 2002. 'The fundamental role of epigenetic events in cancer', *Nat Rev Genet*, 3: 415-28.
Kornberg, T. B., and T. Tabata. 1993. 'Segmentation of the Drosophila embryo', *Curr Opin Genet Dev*, 3: 585-94.
LaFlamme, Brooke. 2014. 'Gene expression in early development', *Nature Genetics*, 46: 99.
Lam, L. L., E. Emberly, H. B. Fraser, S. M. Neumann, E. Chen, G. E. Miller, and M. S. Kobor. 2012. 'Factors underlying variable DNA methylation in a human community cohort', *Proc Natl Acad Sci U S A*, 109 Suppl 2: 17253-60.
Leek, J. T., W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey. 2012. 'The sva package for removing batch effects and other unwanted variation in high-throughput experiments', *Bioinformatics*, 28: 882-3.
Marttila, S., L. Kananen, S. Hayrynen, J. Jylhava, T. Nevalainen, A. Hervonen, M. Jylha, M. Nykter, and M. Hurme. 2015. 'Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression', *BMC Genomics*, 16: 179.
Meinshausen, Nicolai, and Peter Bühlmann. 2010. 'Stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 417-73.

Montgomery, S. B., and E. T. Dermitzakis. 2011. 'From expression QTLs to personalized transcriptomics', *Nat Rev Genet*, 12: 277-82.
Nedelec, Y., J. Sanz, G. Baharian, Z. A. Szpiech, A. Pacis, A. Dumaine, J. C. Grenier, A. Freiman, A. J. Sams, S. Hebert, A. Page Sabourin, F. Luca, R. Blekhman, R. D. Hernandez, R. Pique-Regi, J. Tung, V. Yotova, and L. B. Barreiro. 2016. 'Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens', *Cell*, 167: 657-69 e21.
Orru, V., M. Steri, G. Sole, C. Sidore, F. Virdis, M. Dei, S. Lai, M. Zoledziewska, F. Busonero, A. Mulas, M. Floris, W. I. Mentzen, S. A. Urru, S. Olla, M. Marongiu, M. G. Piras, M. Lobina, A. Maschio, M. Pitzalis, M. F. Urru, M. Marcelli, R. Cusano, F. Deidda, V. Serra, M. Oppo, R. Pilu, F. Reinier, R. Berutti, L. Pireddu, I. Zara, E. Porcu, A. Kwong, C. Brennan, B. Tarrier, R. Lyons, H. M. Kang, S. Uzzau, R. Atzeni, M. Valentini, D. Firinu, L. Leoni, G. Rotta, S. Naitza, A. Angius, M. Congia, M. B. Whalen, C. M. Jones, D. Schlessinger, G. R. Abecasis, E. Fiorillo, S. Sanna, and F. Cucca. 2013. 'Genetic variants regulating immune cell levels in health and disease', *Cell*, 155: 242-56.
Pai, A. A., J. T. Bell, J. C. Marioni, J. K. Pritchard, and Y. Gilad. 2011. 'A genome-wide study of DNA methylation patterns and gene expression levels in multiple human and chimpanzee tissues', *PLoS Genet*, 7: e1001316.
Pai, A. A., J. K. Pritchard, and Y. Gilad. 2015. 'The genetic and mechanistic basis for variation in gene regulation', *PLoS Genet*, 11: e1004857.
Patin, E., J. Bergstedt, S. Ait Kaci Azzou, A. Urrutia, H. Quach, K. Tsuo, L. T. Husquin, M. Rotival, M. S. Kobor, M. L. Albert, D. Duffy, L. Quintana-Murci, and for the Milieu Intérieur Consortium. 2019. 'Factors driving DNA methylation variation in human blood.', *(in preparation)*.
Patin, E., M. Hasan, J. Bergstedt, V. Rouilly, V. Libri, A. Urrutia, C. Alanio, P. Scepanovic, C. Hammer, F. Jonsson, B. Beitz, H. Quach, Y. W. Lim, J. Hunkapiller, M. Zepeda, C. Green, B. Piasecka, C. Leloup, L. Rogge, F. Huetz, I. Peguillet, O. Lantz, M. Fontes, J. P. Di Santo, S. Thomas, J. Fellay, D. Duffy, L. Quintana-Murci, M. L. Albert, and Consortium Milieu Interieur. 2018. 'Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors', *Nat Immunol*, 19: 302-14.
Pearl, J. 2010. 'An introduction to causal inference', *Int J Biostat*, 6: Article 7.
Pellacani, D., M. Bilenky, N. Kannan, A. Heravi-Moussavi, Djhf Knapp, S. Gakkhar, M. Moksa, A. Carles, R. Moore, A. J. Mungall, M. A. Marra, S. J. M. Jones, S. Aparicio, M. Hirst, and C. J. Eaves. 2016. 'Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks', *Cell Rep*, 17: 2060-74.
Peters, M. J., R. Joehanes, L. C. Pilling, C. Schurmann, K. N. Conneely, J. Powell, E. Reinmaa, G. L. Sutphin, A. Zhernakova, K. Schramm, Y. A. Wilson, S. Kobes, T. Tukiainen, Nabec Ukbec Consortium, Y. F. Ramos, H. H. Goring, M. Fornage, Y. Liu, S. A. Gharib, B. E. Stranger, P. L. De Jager, A. Aviv, D. Levy, J. M. Murabito, P. J. Munson, T. Huan, A. Hofman, A. G. Uitterlinden, F. Rivadeneira, J. van Rooij, L. Stolk, L. Broer, M. M. Verbiest, M. Jhamai, P. Arp, A. Metspalu, L. Tserel, L. Milani, N. J. Samani, P. Peterson, S. Kasela, V. Codd, A. Peters, C. K. Ward-Caviness, C. Herder, M. Waldenberger, M. Roden, P. Singmann, S. Zeilinger, T. Illig, G. Homuth, H. J. Grabe, H. Volzke, L. Steil, T. Kocher, A. Murray, D. Melzer, H. Yaghootkar, S. Bandinelli, E. K. Moses, J. W. Kent, J. E. Curran, M. P. Johnson, S. Williams-Blangero, H. J. Westra, A. F. McRae, J. A. Smith, S. L.

Kardia, I. Hovatta, M. Perola, S. Ripatti, V. Salomaa, A. K. Henders, N. G. Martin, A. K. Smith, D. Mehta, E. B. Binder, K. M. Nylocks, E. M. Kennedy, T. Klengel, J. Ding, A. M. Suchy-Dicey, D. A. Enquobahrie, J. Brody, J. I. Rotter, Y. D. Chen, J. Houwing-Duistermaat, M. Kloppenburg, P. E. Slagboom, Q. Helmer, W. den Hollander, S. Bean, T. Raj, N. Bakhshi, Q. P. Wang, L. J. Oyston, B. M. Psaty, R. P. Tracy, G. W. Montgomery, S. T. Turner, J. Blangero, I. Meulenbelt, K. J. Ressler, J. Yang, L. Franke, J. Kettunen, P. M. Visscher, G. G. Neely, R. Korstanje, R. L. Hanson, H. Prokisch, L. Ferrucci, T. Esko, A. Teumer, J. B. van Meurs, and A. D. Johnson. 2015. 'The transcriptional landscape of age in human peripheral blood', *Nat Commun*, 6: 8570.

Piasecka, B., D. Duffy, A. Urrutia, H. Quach, E. Patin, C. Posseme, J. Bergstedt, B. Charbit, V. Rouilly, C. R. MacPherson, M. Hasan, B. Albaud, D. Gentien, J. Fellay, M. L. Albert, L. Quintana-Murci, and Consortium Milieu Interieur. 2018. 'Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges', *Proc Natl Acad Sci U S A*, 115: E488-E97.

Quach, H., M. Rotival, J. Pothlichet, Y. E. Loh, M. Dannemann, N. Zidane, G. Laval, E. Patin, C. Harmant, M. Lopez, M. Deschamps, N. Naffakh, D. Duffy, A. Coen, G. Leroux-Roels, F. Clement, A. Boland, J. F. Deleuze, J. Kelso, M. L. Albert, and L. Quintana-Murci. 2016. 'Genetic Adaptation and Neandertal Admixture Shaped the Immune System of Human Populations', *Cell*, 167: 643-56 e17.

Rahmani, E., R. Schweiger, L. Shenhav, T. Wingert, I. Hofer, E. Gabel, E. Eskin, and E. Halperin. 2018. 'BayesCCE: a Bayesian framework for estimating cell-type composition from DNA methylation without the need for methylation reference', *Genome Biol*, 19: 141.

Raj, A., and A. van Oudenaarden. 2008. 'Nature, nurture, or chance: stochastic gene expression and its consequences', *Cell*, 135: 216-26.

Relton, C. L., and G. Davey Smith. 2012. 'Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease', *Int J Epidemiol*, 41: 161-76.

Richardson, T. G., J. Zheng, G. Davey Smith, N. J. Timpson, T. R. Gaunt, C. L. Relton, and G. Hemani. 2017. 'Mendelian Randomization Analysis Identifies CpG Sites as Putative Mediators for Genetic Influences on Cardiovascular Disease Risk', *Am J Hum Genet*, 101: 590-602.

Roadmap Epigenomics, Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis.

2015. 'Integrative analysis of 111 reference human epigenomes', *Nature*, 518: 317-30.
Roederer, M., L. Quaye, M. Mangino, M. H. Beddall, Y. Mahnke, P. Chattopadhyay, I. Tosi, L. Napolitano, M. Terranova Barberio, C. Menni, F. Villanova, P. Di Meglio, T. D. Spector, and F. O. Nestle. 2015. 'The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis', *Cell*, 161: 387-403.
Shah, Rajen D., and Richard J. Samworth. 2013. 'Variable selection with error control: another look at stability selection', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75: 55-80.
Spurrell, C. H., D. E. Dickel, and A. Visel. 2016. 'The Ties That Bind: Mapping the Dynamic Enhancer-Promoter Interactome', *Cell*, 167: 1163-66.
Sugawara, H., K. Iwamoto, M. Bundo, J. Ueda, J. Ishigooka, and T. Kato. 2011. 'Comprehensive DNA methylation analysis of human peripheral blood leukocytes and lymphoblastoid cell lines', *Epigenetics*, 6: 508-15.
Tedder, T. F., L. J. Zhou, and P. Engel. 1994. 'The CD19/CD21 signal transduction complex of B lymphocytes', *Immunol Today*, 15: 437-42.
Teschendorff, A. E., and C. L. Relton. 2018. 'Statistical and integrative system-level analysis of DNA methylation data', *Nat Rev Genet*, 19: 129-47.
Thomas, S., V. Rouilly, E. Patin, C. Alanio, A. Dubois, C. Delval, L. G. Marquier, N. Fauchoux, S. Sayegrih, M. Vray, D. Duffy, L. Quintana-Murci, M. L. Albert, and Consortium Milieu Interieur. 2015. 'The Milieu Interieur study - an integrative approach for study of human immunological variance', *Clin Immunol*, 157: 277-93.
Tung, J., and Y. Gilad. 2013. 'Social environmental effects on gene regulation', *Cell Mol Life Sci*, 70: 4323-39.
Urrutia, A., D. Duffy, V. Rouilly, C. Posseme, R. Djebali, G. Illanes, V. Libri, B. Albaud, D. Gentien, B. Piasecka, M. Hasan, M. Fontes, L. Quintana-Murci, M. L. Albert, and Consortium Milieu Interieur. 2016. 'Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses', *Cell Rep*, 16: 2777-91.
Vogel, L. B., and D. J. Fujita. 1995. 'p70 phosphorylation and binding to p56lck is an early event in interleukin-2-induced onset of cell cycle progression in T-lymphocytes', *J Biol Chem*, 270: 2506-11.
Võsa, Urmo, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Silva Kasela, Natalia Pervjakova, Isabel Alvaes, Marie-Julie Fave, Mawusse Agbessi, Mark Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg-Guzman, Johannes Kettunen, Joseph Powell, Bernett Lee, Futao Zhang, Wibowo Arindrarto, Frank Beutner, Harm Brugge, Julia Dmitreva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T. Heijmans, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M. Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel Nivard, Brenda Penninx, Jonathan Pritchard, Olli Raitakari, Olaf Rotzchke, Eline P. Slagboom, Coen D. A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A. C. Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten

van Iterson, Jan Veldink, Uwe Völker, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann, Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon Pierce, Terho Lehtimäki, Dorret Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Jian Yang, Peter M. Visscher, Markus Scholz, Gregory Gibson, Tõnu Esko, and Lude Franke. 2018. 'Unraveling the polygenic architecture of complex traits using blood eQTL metaanalysis', *bioRxiv*: 447367.

Wagner, J. R., S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette. 2014. 'The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts', *Genome Biol*, 15: R37.

Willinger, T., T. Freeman, M. Herbert, H. Hasegawa, A. J. McMichael, and M. F. Callan. 2006. 'Human naive CD8 T cells down-regulate expression of the WNT pathway transcription factors lymphoid enhancer binding factor 1 and transcription factor 7 (T cell factor-1) following antigen encounter in vitro and in vivo', *J Immunol*, 176: 1439-46.

Yang, X., H. Han, D. D. De Carvalho, F. D. Lay, P. A. Jones, and G. Liang. 2014. 'Gene body methylation can alter gene expression and is a therapeutic target in cancer', *Cancer Cell*, 26: 577-90.

Yevshin, I., R. Sharipov, S. Kolmykov, Y. Kondrakhin, and F. Kolpakov. 2019. 'GTRD: a database on gene transcription regulation-2019 update', *Nucleic Acids Res*, 47: D100-D05.

Yin, Y., E. Morgunova, A. Jolma, E. Kaasinen, B. Sahu, S. Khund-Sayeed, P. K. Das, T. Kivioja, K. Dave, F. Zhong, K. R. Nitta, M. Taipale, A. Popov, P. A. Ginno, S. Domcke, J. Yan, D. Schubeler, C. Vinson, and J. Taipale. 2017. 'Impact of cytosine methylation on DNA binding specificities of human transcription factors', *Science*, 356.

Zheng, S. C., C. E. Breeze, S. Beck, and A. E. Teschendorff. 2018. 'Identification of differentially methylated cell types in epigenome-wide association studies', *Nat Methods*, 15: 1059-66.

Zhou, J., R. L. Sears, X. Xing, B. Zhang, D. Li, N. B. Rockweiler, H. S. Jang, M. N. K. Choudhary, H. J. Lee, R. F. Lowdon, J. Arand, B. Tabers, C. C. Gu, T. J. Cicero, and T. Wang. 2017. 'Tissue-specific DNA methylation is conserved across human, mouse, and rat, and driven by primary sequence conservation', *BMC Genomics*, 18: 724.

Zhu, H., G. Wang, and J. Qian. 2016. 'Transcription factors as readers and effectors of DNA methylation', *Nat Rev Genet*, 17: 551-65.

Supplementary Figures

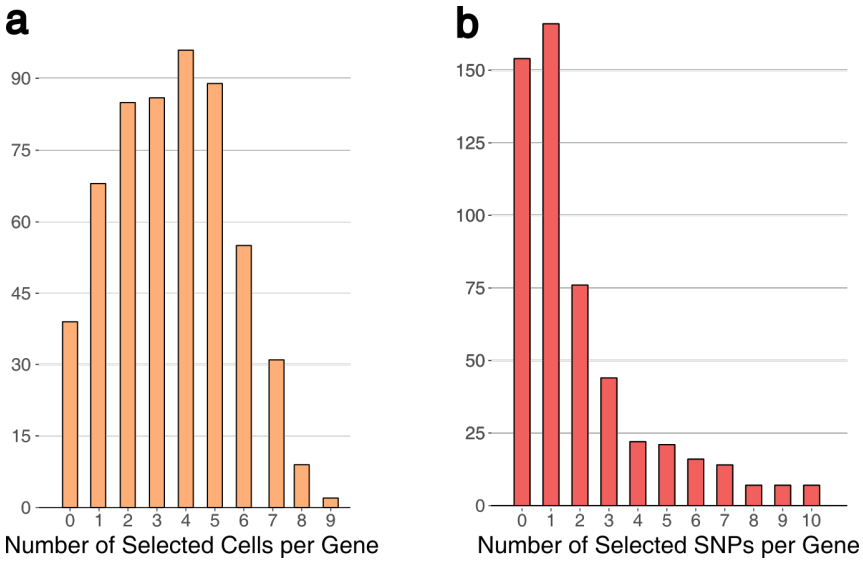

Figure S1
a, Number of genes, on the y-axis, as a function of the number of cell-types that were selected with the stability selection algorithm, on the x-axis. **b,** Number of genes, on the y-axis, as a function of the number of SNPs that were selected with the stability selection algorithm, on the x-axis.

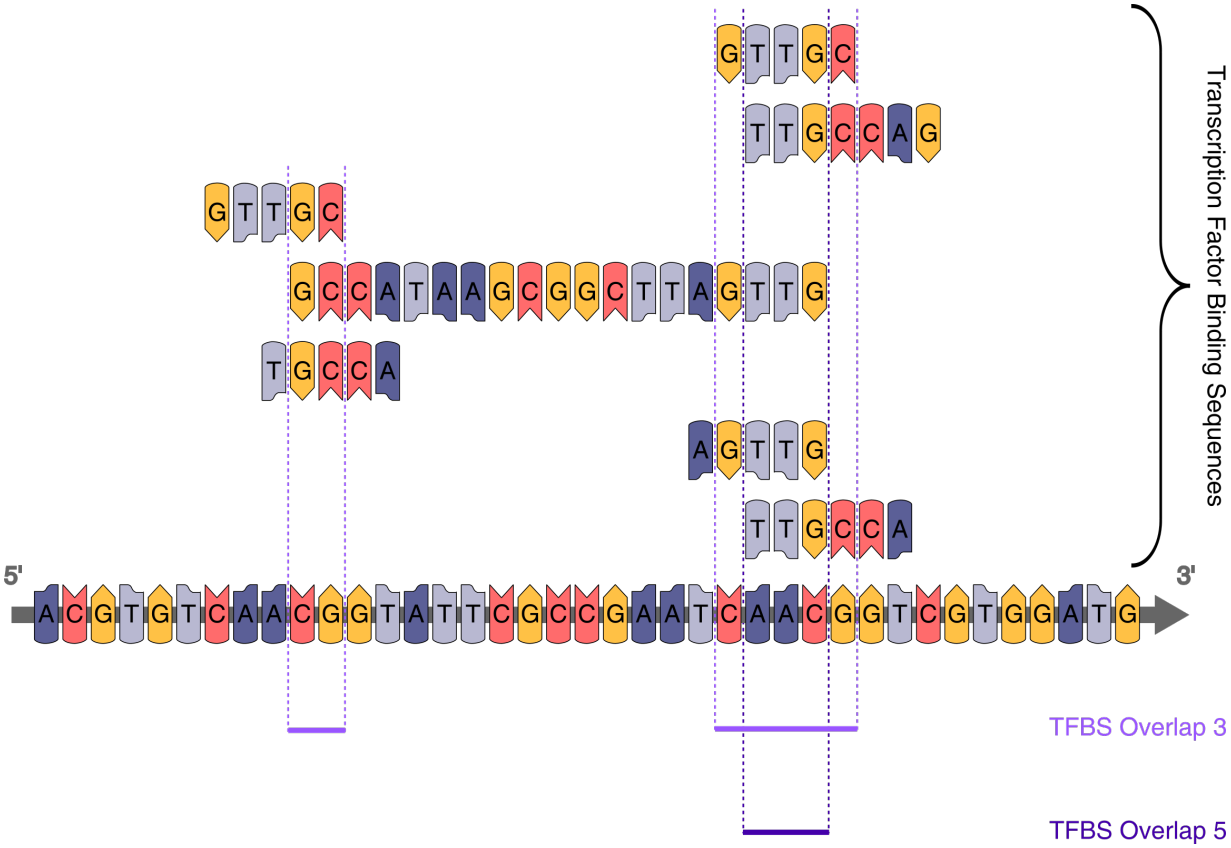

Figure S2
Rationale for the detection of TFBS overlaps regions. An schematic example of a genomic region with two TFBS overlap of 3, and one TFBS overlap of 5.

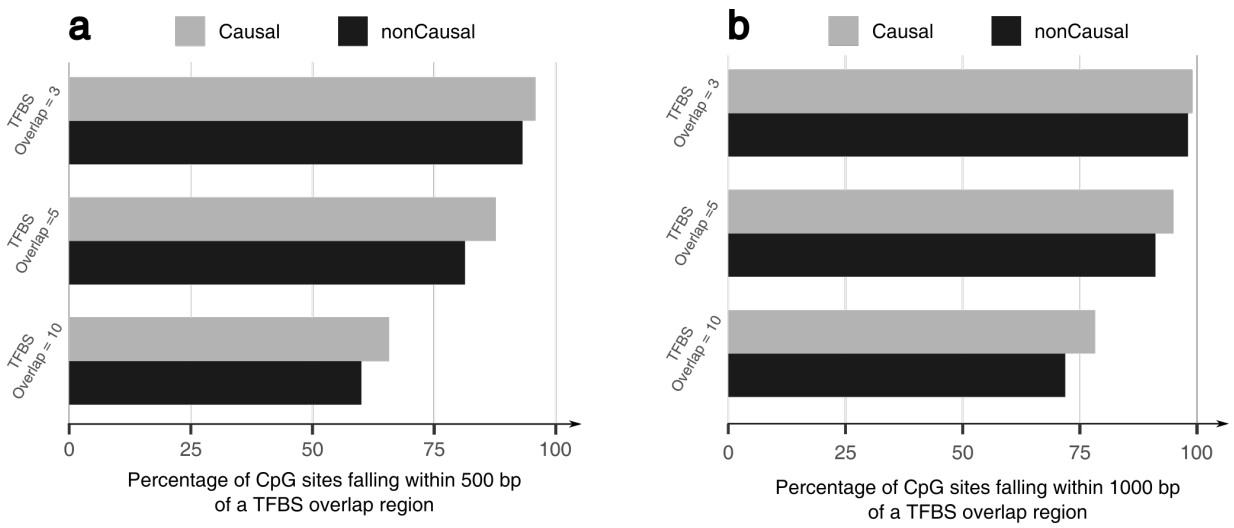

Figure S3

a, Enrichment of eQTMs in TFBS overlap regions. The proportions of eQTMs falling within 500 bp of the center of a region where at least 3, 5 or 10 TFBS overlapped is displayed in grey bars. The black bars represent the proportions of the CpGs that were not detected as eQTMs among the 52,614 test CpGs. **b,** Enrichment of eQTMs in TFBS overlap regions. The proportions of eQTMs falling within 1,000 bp of the center of a region where at least 3, 5 or 10 TFBS overlapped is displayed in grey bars. The black bars represent the proportions of the CpGs that were not detected as eQTMs among the 52,614 test CpGs.

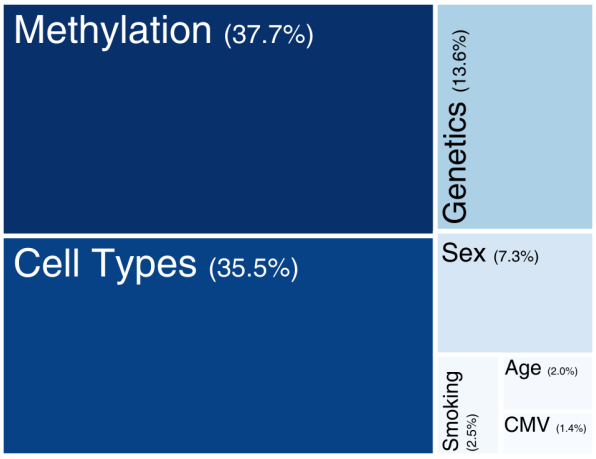

Figure S4
Treemap plot of the sum across all genes for which we detected at least one significant eQTM of the variance explained by the 7 main different variables included in the model: Methylation, Cell-types, Genetics, Sex, Smoking, Age and CMV seropositivity. The area and shade of blue of each rectangle is proportional to the value indicated between parentheses, which represent the overall impact of each variable on gene expression variance.

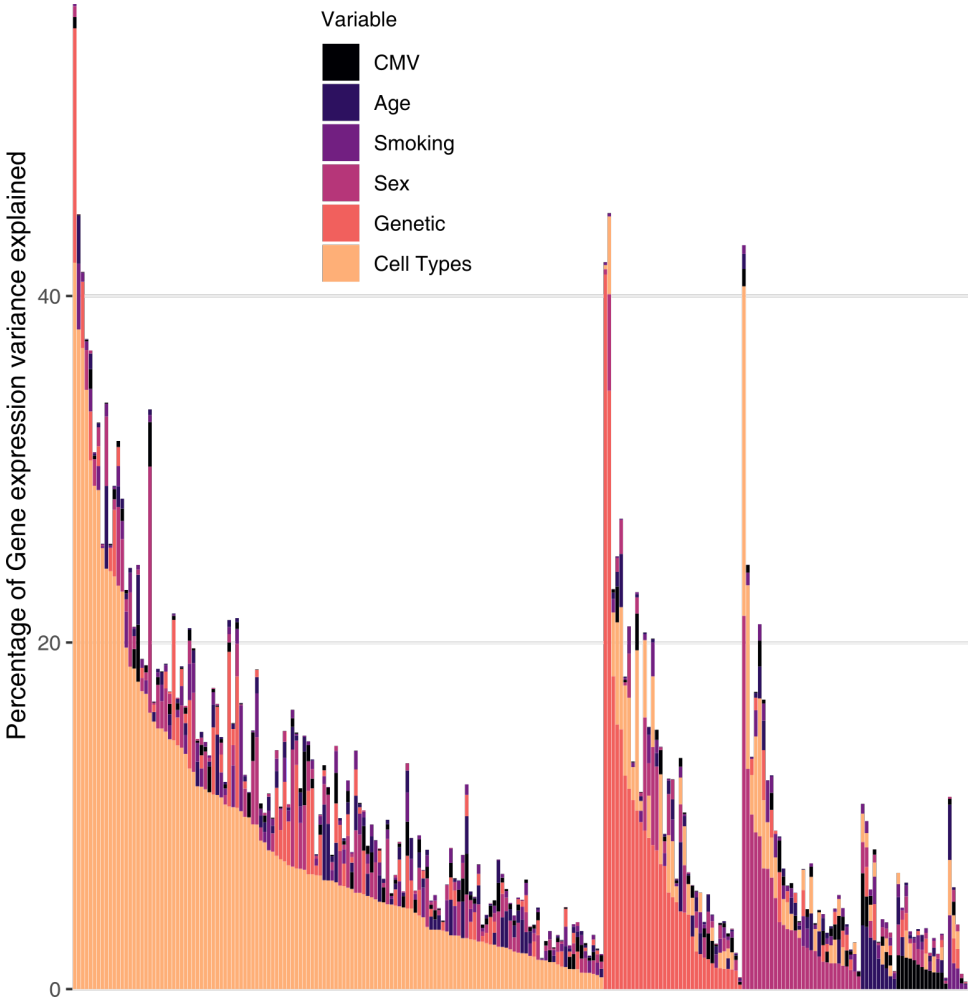

Figure S5
Bar plot of the percentages of variance explained by the different variables included in the model for each of the 176 gene for which we could not find any significant eQTM. Each line of the plot represent the total percentage of variance explained for a given gene, with the different colors indicating the respective impact of each variable. Genes are ordered according to the main variable explaining the most amount of variance of their expression.

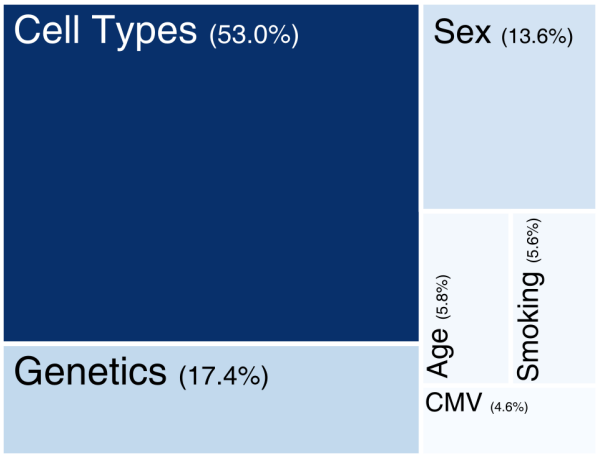

Figure S6
Treemap plot of the sum across all 176 genes for which we could not find any significant eQTM of the variance explained by the 6 main different variables included in the model: Cell-types, Genetics, Sex, Smoking, Age and CMV seropositivity. The area and shade of blue of each rectangle is proportional to the value indicated between parentheses, which represent the overall impact of each variable on gene expression variance.

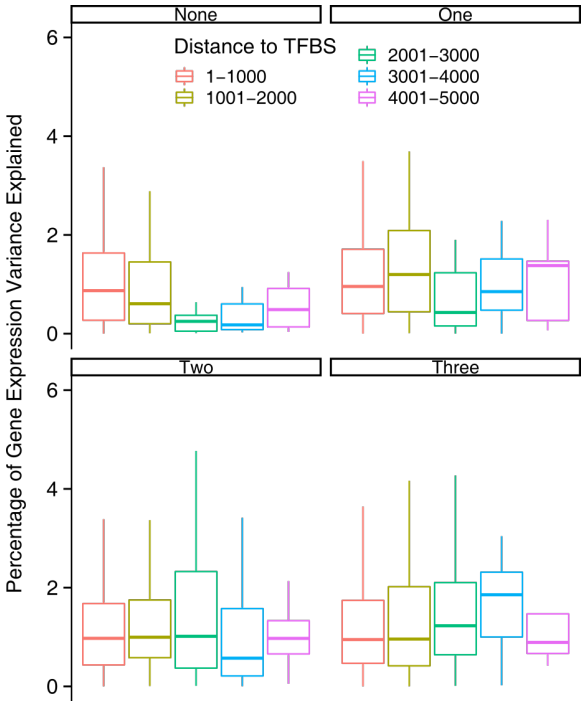

Figure S7
Boxplots displaying the percentage of variance explained by the different eQTMs, according to their distance to the nearest TFBS overlap region of at least 10 TFBS. eQTMs were separated in 4 categories, according to whether they fell in 0, 1, 2 or 3 distinct regulatory regions (among PIRs, enhancer and promoter regions).

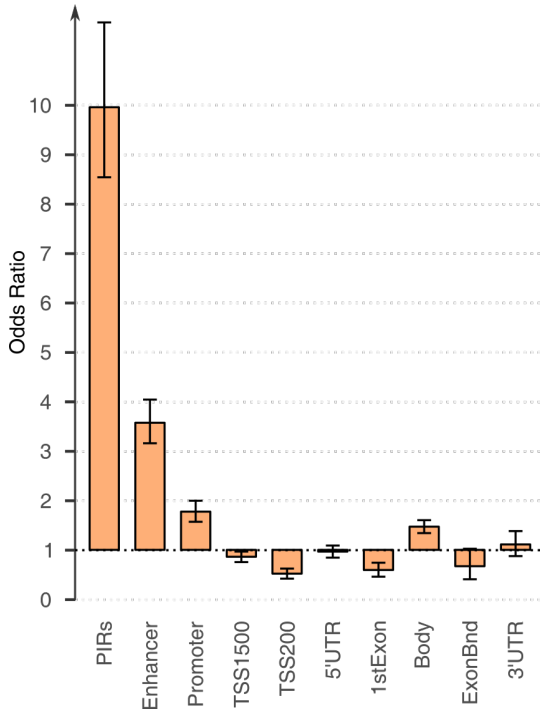

Figure S8
Genomic location of cell-eQTMs, in light orange. Odds ratio and 95% confidence intervals are displayed for the 1,144 cell-eQTMs detected in this study, comparing their localization in 7 genomic locations provided by Illumina (TSS1500, TSS200, 5'UTR, 1stExon, Body, ExonBnd and 3'UTR), and in enhancer, promoter and PIRs. Odds ratio were computed against the distribution of the 52,614 tested CpGs.

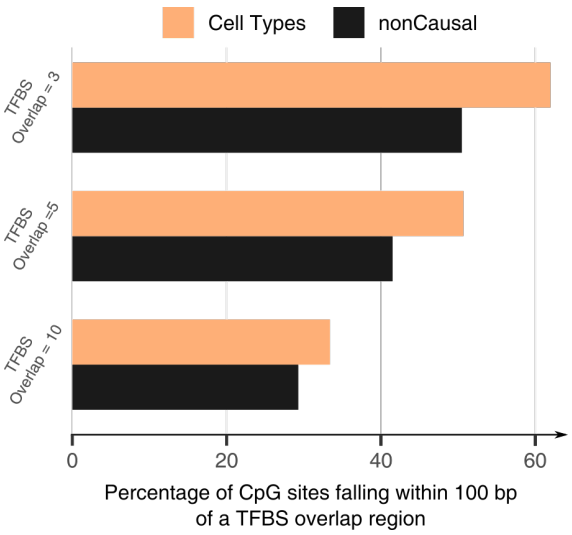

Figure S9
Enrichment of cell-eQTMs in TFBS overlap regions. The proportions of cell-eQTMs falling within 100 bp of the center of a region where at least 3, 5 or 10 TFBS overlapped is displayed in light orange bars. The black bars represent the proportions of the 52,614 tested CpGs.

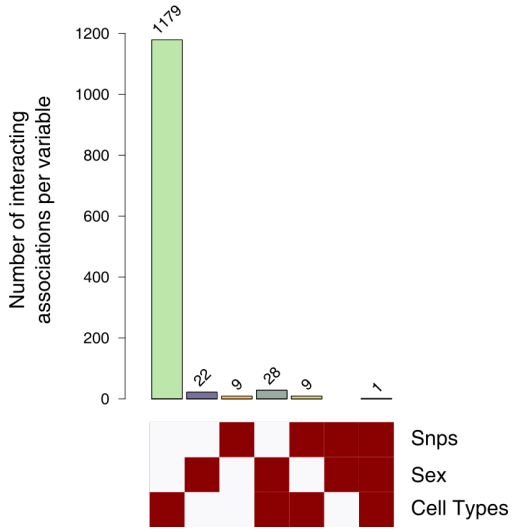

Figure S10
Number of CpG-gene associations with a significant interaction effect with one, two or three variables among cell types, sex and SNPs.

6.3 Résumé des résultats et perspectives

L'utilisation d'un graphe de causalité nous a permis de théoriser l'ensemble de nos a priori vis à vis des variables pouvant potentiellement affecter la variabilité de la méthylation de l'ADN ou de l'expression. Dans le cadre de ce graphe, nous avons identifié 2,939 associations significatives entre variabilité de la méthylation et niveau d'expression des gènes de l'immunité. Parmi les 542 gènes testés, nous avons pu détecté au moins une association significative pour 334 gènes, indiquant que l'expression d'environ 62% des gènes de l'immunité est régulée de manière causale par la méthylation d'au moins un *cis*-eQTM. Puisque nous avons utilisé pour chaque gène les types cellulaires qui leur étaient spécifiquement associés comme variables d'ajustement, nos résultats ne sont pas déterminés par des effets dus à l'hétérogénéité cellulaire du sang. Au contraire, nous avons encore approfondi notre étude en ajoutant un terme d'interaction entre la méthylation et les proportions des différents types cellulaires dans le modèle expliquant l'expression des gènes. Ceci nous a permis d'identifier les eQTM pour lesquels l'effet de la méthylation sur l'expression était significativement dépendant des proportions en certains types cellulaires (cell-eQTM). Remarquablement, 1,217 associations (41.4% de toutes les associations détectées) ont montré un tel effet cellule-spécifique, soulignant l'importance d'ajuster pour l'hétérogénéité en types cellulaires dans de telles études d'association.

Nous nous sommes ensuite intéressés à l'emplacement des eQTM et cell-eQTM le long du génome, relativement à la répartition de l'ensemble des sites CpG que nous avions testés. Ceci nous a permis d'identifier les régions génomiques enrichies en méthylation causalement impliquée dans la régulation de l'expression. Pour ce faire, nous avons utilisé différentes bases de données des régions régulatrices de l'expression, telles que les promoteurs, les *enhancers*, les PIR ou les sites de fixation de facteurs de transcription. Nos résultats nous ont permis de confirmer l'importance de ces régions dans la régulation de l'expression puisque nous avons trouvé un fort enrichissement dans chacune de ces régions (ratios entre 1.2 et 2.7 pour les eQTM, et entre 1.8 et 10.0 pour les cell-eQTM). De tels niveaux d'enrichissement apportent une nouvelle preuve quant à l'importance de ces régions, mais surtout nous permettent d'émettre l'hypothèse que les eQTM ne tombant dans aucune des susdites régions sont potentiellement des marqueurs de régions n'ayant pas encore été découvertes. Ces résultats sont quoiqu'il en soit de bons arguments quant à l'inclusion des profils de méthylation dans les analyses de détection de telles régions régulatrices. Enfin, la comparaison de la localisation des cell-eQTM à celle du reste des eQTM nous a permis d'identifier un très fort enrichissement dans les PIR (ratio de 4.8), ce qui peut suggérer que ces régions sont potentiellement impliquées dans la régulation de l'expression des gènes montrant des patrons d'expression cellule-spécifique, en opposition aux autres régions régulatrices, comme les *enhancers* par exemple, qui seraient plus généralement impliquées dans la régulation des gènes communs à toutes les cellules sanguines.

Finalement, malgré le fort lien existant entre l'âge et la méthylation de certains sites CpG, que nous avons discuté dans les chapitres précédents, au travers de l'exemple des horloges épigénétiques en particulier, nous n'avons trouvé aucun effet de l'âge sur les associations entre méthylation et expression que nous avons détectées. Ce résultat surprenant semble indiquer que les sites CpG dont le niveau de méthylation varie au cours de la vie n'ont peu ou pas d'impact sur le niveau d'expression des gènes attenants dans les cellules du sang, à l'exception peut-être des cellules cancéreuses.

Chapitre 7

Discussion générale et perspectives

7.1 Vers une compréhension plus globale des mécanismes de régulation de l'expression des gènes et de leurs conséquences

7.1.1 L'impact de la méthylation dans les différences transcriptionnelles inter-populationnelles

Comme nous l'avons précédemment abordé, il existe un grand nombre d'études décrivant les différences d'expression entre populations, ainsi que les bases génétiques associées à de telles différences (Spielman et al., 2007 ; W. Zhang et al., 2008 ; Stranger et al., 2012 ; Quach et al., 2016 ; Nedelec et al., 2016 ; Kim-Hellmuth et al., 2017). Puisque ces différences d'expression peuvent, en partie, nous permettre de comprendre les différences phénotypiques, dans notre capacité à résister aux pathogènes par exemple, il y a un intérêt certain à identifier les causes de ces différences inter-populationnelles. Or, puisque la méthylation de l'expression peut apporter un niveau d'information supplémentaire dans le contrôle de la transcription, l'analyse des différences de méthylation entre populations, et de leur rôle causal dans la régulation de l'expression, a le potentiel d'expliquer une partie des différences inter-populationnelles que la génétique seule n'a pu résoudre.

Plusieurs résultats présentés dans le chapitre 5 vont dans ce sens. Tout d'abord, l'importante différence de méthylation que nous avons observée entre individus d'origines européenne et africaine, confirme les résultats des autres analyses ayant étudié les différences inter-populationnelles de la méthylation de l'ADN (Fraser et al., 2012 ; Moen et al., 2013 ; Heyn et al., 2013 ; Carja et al., 2017). Par ailleurs, nous avons identifié que les sites CpG différentiellement méthylés étaient enrichis en eQTM, indiquant qu'une partie des différences d'expression entre populations pouvait être expliquée par des variations inter-populationnelles du niveau de méthylation de certains CpG, en accord avec une étude datant de 2013 (Moen et al., 2013). Enfin, nous avons détecté un plus haut niveau de méthylation dans les individus d'origine africaine, relativement aux individus d'origine européenne, un résultat qui s'inscrit à contre-courant d'un grand nombre d'études sur les différences de méthylation entre Africains et Européens, dont les résultats sont résumés dans cette revue (Kader & Ghai, 2017). De plus, la forte dissymétrie (75% des DMS sont hyper-méthylés chez les Africains) que nous avons observée a une composante génétique puisque les allèles associés à une plus forte méthylation sont en général plus fréquents chez les individus d'origine africaine. Ceci résulte donc en une méthylation plus élevée

chez ces individus, et suggère potentiellement un phénomène d'adaptation, une hypothèse qui nécessite toutefois d'être validée par d'autres études.

7.1.2 Apport des études de causalité

Le développement des méthodes de séquençage ces dernières années a conduit à une disponibilité croissante de jeux de données regroupant profilages génétique, épigénétique et transcriptomiques pour les mêmes individus, permettant ainsi d'explorer les bases génétiques et épigénétiques de la régulation de l'expression (Wagner et al., 2014 ; Bonder et al., 2014 ; Gutierrez-Arcelus et al., 2015 ; Bonder et al., 2017). Ce type d'études est venu compléter notre compréhension des effets que les facteurs génétiques et épigénétiques avaient indépendemment sur l'expression des gènes. En particulier, leur apport majeur a été de révéler la forte concomitance des variants génétiques et épigénétiques dans les processus de régulation de la transcription (Bonder et al., 2014, 2017 ; Husquin et al., 2018). Il apparaît en effet qu'une énorme majorité des eQTM est aussi sous contrôle génétique, jusqu'à 80% comme nous l'avons identifié dans le chapitre 5, ce qui met en doute le réel impact de la méthylation dans la régulation de l'expression (figure 3.1, page 39).

Dans la perspective d'apporter des réponses à cette incertitude, l'utilisation de différentes méthodes d'inférences, telles que les méthodes de médiation, de corrélations partielles, ou de randomisation mendélienne, a permis de déterminer le véritable impact de la méthylation sur l'expression des gènes (van Eijk et al., 2012 ; Gutierrez-Arcelus et al., 2013 ; Husquin et al., 2018). Ces études ont donc permis de focaliser les résultats de recherche d'eQTM sur les sites CpG qui ont un impact causal sur la transcription des gènes, et ce faisant, de mieux caractériser les régions du génome impliquées dans la régulation de l'expression. Ainsi, nous avons pu montrer dans le chapitre 5 que 86.6% des sites CpG identifiés via notre analyse de médiation sont localisés dans les sites de fixation des facteurs de transcription (TF). Cet enrichissement (1.9 fois plus qu'attendu) pourrait permettre d'établir la liste des TF spécifiques aux monocytes. En effet, plusieurs études ont montré qu'il était possible d'identifier les TF spécifiques à un tissu en observant les profils de méthylation des sites de fixation des TF dans ce même tissu (Ziller et al., 2013 ; Yuan et al., 2015). Par ailleurs, les analyses de médiation sont un outil intéressant afin de déterminer les effets fonctionnels des variants génétiques détectés par GWAS, en permettant de lier la variabilité génétique de ces SNPs à l'expression des gènes au travers de la variabilité épigénétique. Comme montré dans le chapitre 5, il existe en effet un fort enrichissement des meQTL en variants génétiques précédemment identifiés comme résultats de GWAS. Il est certain que dans un certain nombre de situations, l'effet des facteurs génétiques de risques ont un effet sur le phénotype via l'effet de la méthylation, comme dans le cas de la polyarthrite rhumatoïde (Liu et al., 2013).

Cependant, l'étude de la causalité ne va pas sans un certain nombre de défis, à commencer par la grande sensibilité des méthodes d'inférences de causalité aux erreurs de mesures. De plus, les conclusions de ces analyses sont peu fiables si les hypothèses initiales — telles que la linéarité — qui peuvent se révéler compliquées à tester, ne sont pas respectées (Teschendorff & Relton, 2018). Afin d'améliorer l'inférence des relations de cause à effet, il a été récemment proposé d'intégrer des a priori biologiques aux modèles de causalité (Lappalainen & Greally, 2017). C'est dans cette optique que nous avons conçu l'analyse de causalité décrite dans le chapitre 6. En effet, nous avons utilisé l'ensemble

des données à notre disposition afin de construire un graphe de causalité qui décrit nos a priori des relations de causes à effets existant entre les différentes variables du système. Ce faisant, nous avons pu appliquer un modèle linéaire pour expliquer la variabilité de l'expression au regard des interactions décrites dans ce graphe, et ainsi, quantifier l'effet causal de la méthylation de l'ADN. Par ailleurs, l'intégration d'un nombre croissant de différents types de données devrait permettre d'affiner notre compréhension des mécanismes de régulation de l'expression des gènes. L'intérêt d'une telle approche a été montrée dans une étude combinant des données génétiques et d'expression avec des données de trois marques épigénétiques (H3K27ac, H3K4me1 et H3K4me3), connues pour être des marqueurs de l'activité des *enhancers* et des promoteurs (Delaneau et al., 2019). Dans un premier temps, les auteurs ont utilisé la forte corrélation locale de ces marques épigénétiques pour identifier des domaines de *cis*-régulation (*Cis-Regulatory Domains*, CRD) et se sont aperçus qu'il existait un très grand nombre d'interactions physiques entre ces CRD, que ce soit au sein d'un même chromosome ou entre chromosomes. Finalement, en utilisant les données génétiques et d'expression, les auteurs ont détecté différents moyens par lesquels l'effet de la génétique sur l'expression était médié par l'activité de ces CRD, apportant une preuve du rôle causal sur l'expression que peut avoir l'action coordonnée de différentes marques épigénétiques le long du génome.

7.1.3 Pour aller plus loin : apport des séries temporelles

Nous venons de voir comment la multiplication des types de données peut permettre d'apporter une réponse aux questions de causalité entre génétique, épigénétique et expression des gènes. Une autre approche repose sur la multiplication, non pas des types de données mais, en se concentrant sur une seule marque épigénétique, des données chez les mêmes individus de manière répétée au cours du temps. L'intérêt immédiat de ce genre d'approche est de répondre à l'équivalent du paradoxe populaire "Qui de l'oeuf ou la poule est apparu en premier ?", appliqué à la méthylation et l'expression de l'ADN. En effet, il est admis que les deux phénomènes — la méthylation affecte l'expression, et l'expression affecte la méthylation — se produisent, et en conséquence, un débat existe pour savoir qui de l'expression ou de la méthylation est affecté en premier en réponse à une modification de l'environnement.

Dans les résultats présentés chapitre 5 et 6, nous avons considéré les phénomènes de causalité inversée (*reverse causation*) comme hautement improbables en conséquence du timing particulier de l'acquisition des données dans ces deux projets. En effet, dans les deux cas, après purification des monocytes (chapitre 5) ou obtention du sang (chapitre 6), l'ADN a été extrait à t0, alors que l'extraction des ARNm n'a été conduite qu'après 6h. Ainsi, pour que les profils d'expression que nous avons analysés aient un impact causal sur la méthylation cela nécessiterait une très forte corrélation entre l'expression à t0 et à t0+6h, un postulat que nous avons considéré comme improbable. Toutefois, si cette différence de timing nous a donc permis de nous concentrer sur les liens de cause à effet de la méthylation de l'ADN vers l'expression uniquement, elle n'en reste pas moins un biais majeur de nos résultats. En conséquence, une manière plus élégante de faire face à cette question de causalité entre ces deux variables eût été d'extraire l'ADN et l'ARN de manière concomitante, à plusieurs intervalles de temps après stimulation. Un tel mode opératoire nous aurait alors permis d'étudier en détail les dynamiques des profils

de méthylation et d'expression au cours de la réponse des monocytes à l'activation des différentes voies immunitaires stimulées.

C'est dans cette optique qu'une étude sur la réponse des cellules dendritiques à l'infection par *Mycobacterium tuberculosis* a été menée. Nous avons discuté au cours du chapitre 3 des résultats de l'étude initiale qui avait détecté une déméthylation rapide et active de plusieurs milliers de sites CpG suite à l'infection (Pacis et al., 2015). Toutefois, si l'étude initiale s'était concentrée sur la méthylation et l'expression en un seul point dans le temps après l'infection, les auteurs ont complété leurs premiers résultats en répétant leur expérience pour mesurer la méthylation et l'expression à t=2, 18, 48, et 72 h post-infection (Pacis et al., 2019). Ce faisant, ils ont identifié les gènes différentiellement exprimés (DE) et les sites CpG différentiellement méthylés (DM) en réponse à l'infection. En comparant les DE avec les DM, ils ont détecté que pour 83.1% des gènes associés avec une hausse de l'expression et une baisse de la méthylation, les changements de l'expression précédaient les changements de méthylation, l'inverse n'étant vrai que pour 8 gènes (1.3%). Enfin, les auteurs ont déterminé que les changements de méthylation étaient en majeure partie localisés dans les *enhancers*, tandis que les promoteurs semblaient réfractaires aux modifications ; et que ces changements succédaient à la fixation des TF et à la modification d'autre marques épigénétiques, telles que l'acéthylation de la lysine 27 de l'histone H3 (H3K27ac). En conclusion, les auteurs arguent que la méthylation de l'ADN ne joue donc pas un rôle primordial dans la régulation des mécanismes des premières phases de la réponse immunitaire innée. Cependant, puisque les modifications de méthylation persistent 3 jours après l'infection, ils émettent l'hypothèse que ces modifications pourraient jouer un rôle dans les phases terminales de la réponse immunitaire et/ou être impliquées dans l'entraînement de l'immunité innée, en permettant une réponse plus rapide à des infections futures par des pathogènes similaires. Notons toutefois quelques limitations à cette étude : tout d'abord, les associations faites entre CpG et gènes sont basées sur la colocalisation de différences d'expression et de méthylation. Comme les *enhancers* peuvent contrôler l'expression des gènes tout en étant situés à de grandes distances des gènes, une partie des associations entre sites CpG et gènes est donc négligée. De plus, et c'est le biais principal de cette étude, aucun test d'association n'a été conduit pour identifier les relations causales entre méthylation et expression. De ce fait, je suis convaincu que l'utilisation de méthodes d'inférences de causalité telles que décrites précédemment et l'intégration de données génétiques à ce type d'étude permettraient de mieux distinguer les directions des relations de cause à effet entre méthylation et expression à l'échelle du génome, sans se limiter aux associations locales.

Un autre aspect de cette idée d'observer la dynamique de la méthylation et de l'expression au cours du temps serait de se placer à l'échelle inter-générationnelle, en observant les profils d'expression et de méthylation dans le même tissu au cours de générations successives. Ce genre d'étude, menée chez des organismes à faible temps de génération, tels que les bactéries, devrait permettre de répondre à des questions reliées à la transmission inter- et trans-générationnelle des profils de méthylation et de leur rôle dans les phénomènes d'adaptation, comme proposé par Klironomos (Klironomos et al., 2013).

7.2 Vers une synthèse inclusive de la théorie de l'évolution

Bien que la découverte de l'ADN et l'essor de la génétique nous aient permis d'améliorer considérablement notre compréhension de l'hérédité et de la biologie évolutive, on peut aussi y voir des répercussions négatives. En particulier, l'un des principaux contrecoups a été de restreindre notre représentation des phénomènes d'hérédité et d'évolution, avec pour conséquence directe de nous désintéresser des mécanismes héréditaires ne reposant pas sur la transmission d'une séquence d'ADN. Dans cette partie, nous allons tenter d'enlever ces œillères pour présenter différents arguments en faveur d'une réécriture de la théorie de l'évolution telle qu'elle est conçue de nos jours.

7.2.1 Comment l'expression des gènes peut contribuer à l'adaptation locale

Commençons tout d'abord par nous intéresser à la signification évolutive du terme "adaptation", en le distinguant du sens commun. La définition principale du larousse est la suivante "Action d'adapter ou de s'adapter à quelque chose : Adaptation aux circonstances.", tandis qu'une définition admise dans la biologie de l'évolution peut être la suivante : "Les processus par lesquels la sélection naturelle affecte de manière inclusive les variations héritables d'une génération à l'autre, de façon à augmenter la chance de survie et de reproduction des organismes dans leur environnement." (Danchin, Pocheville, & Huneman, 2019). Il est donc nécessaire de distinguer l'adaptation des individus à leur environnement, au cours de leur vie, de l'adaptation dans le sens de la sélection naturelle. Pour cela définissons l'acclimatation (*accomodation*, en anglais) par "les processus par lesquels les organismes répondent à des changements de leur environnement en modifiant leur phénotype. Cela permet aux organismes d'améliorer leurs chances de survie et de reproduction dans leur environnement actuel, mais ces processus ne sont pas transmis d'une génération à l'autre." (Danchin, Pocheville, & Huneman, 2019).

Maintenant que nous avons défini ces deux termes, discutons de la contribution de la diversité des profils d'expression dans les phénomènes d'adaptation. L'idée que les régions régulatrices de l'expression puissent être sélectionnées est vieille de plusieurs décennies : après leur découverte de l'opéron lactose en 1961, Jacob et Monod ont très rapidement émis l'hypothèse que les régions régulatrices jouaient un rôle unique durant l'évolution. Il a néanmoins fallu attendre le début des années 2000 et l'émergence des données génétiques pour qu'il devienne évident aujourd'hui que les mutations au sein des régions régulatrices contribuent en effet à des différences phénotypiques adaptatives (Wray, 2007). Depuis ces premières études, il a été largement démontré que les phénomènes d'adaptation locale (qui ne sont retrouvés que chez certaines populations humaines) étaient bien plus fréquemment dues à des modifications des profils d'expression, plutôt qu'à des modifications des séquences codantes (Fraser, 2013). Plus récemment encore, une comparaison du transcriptome des variétés cultivées de tomate, du maïs et de haricot à celui de leur ancêtre sauvage a mis en évidence des changements d'expression de plusieurs milliers de gènes, ce qui suggère que l'expression des gènes a probablement été ciblée par la sélection artificielle au cours de la domestication des plantes alimentaires (Hufford et al., 2012 ; Swanson-Wagner et al., 2012 ; Koenig et al., 2013 ; Bellucci et al., 2014). Cette implication

de l'expression des gènes dans la réponse adaptative à des pressions environnementales a aussi été identifiée chez l'espèce humaine, en particulier dans notre lutte contre les pathogènes (Quach et al., 2016 ; Nedelec et al., 2016 ; Harrison et al., 2019). De plus, un certain nombre de résultats pointent vers le potentiel adaptatif de la diversité des profils d'expression, qui permettrait de faire face à une plus grande variété de l'environnement. En particulier, une étude sur les profils d'expression de *Miscanthus lutarioriparius*, une variété de plante herbacée, a montré une augmentation de la diversité de l'expression en même temps qu'une diminution de la diversité génétique, en conséquence d'un changement d'environnement (Xu et al., 2015). En outre, les auteurs ont identifié que les fonctions des gènes dont la variabilité de l'expression avait le plus augmenté sont les plus pertinentes vis à vis des caractéristiques du nouvel environnement. Ce résultat suggère donc qu'une plus grande diversité de l'expression pourrait représenter un intérêt adaptatif, en offrant une plus grande variabilité phénotypique sur laquelle la sélection naturelle peut agir.

Toutefois, comme nous l'avons rappelé dans la définition de l'adaptation ci-dessus, la sélection naturelle doit agir sur "les variations héritables d'une génération à l'autre", un raisonnement contre-intuitif ici, si l'on considère que les variations héritables sont uniquement celles affectant les séquences d'ADN. En effet, dans l'exemple de *Miscanthus lutarioriparius*, la diversité de l'expression, et en conséquence phénotypique, augmente tandis que la diversité génétique diminue. Or, si l'on considère la plasticité phénotypique comme un processus adaptatif, il faut identifier les "variants héritables" responsables de cette plasticité, c'est à dire les causes de l'augmentation de la variabilité de l'expression. Comme il semble improbable qu'une diminution de la diversité génétique soit responsable d'une augmentation de la diversité de l'expression, il est nécessaire de chercher à un autre niveau. Ici rentre en jeu la variabilité épigénétique, dont nous allons maintenant discuter comment elle peut affecter le tempo et l'issue des phénomènes d'adaptation.

7.2.2 Le rôle de la méthylation de l'ADN dans la plasticité phénotypique

Commençons par rappeler le rôle de la méthylation dans l'établissement des phénotypes cellulaires spécifiques, que nous avons décrit dans les chapitres précédents. En particulier, nos résultats du chapitre 6 indiquent que l'expression de près de 2 gènes de l'immunité sur 3 est régulée de manière causale par la méthylation dans les cellules sanguines. De plus, nous avons mis en évidence qu'un très grand nombre d'associations présentaient des effets cellule-spécifiques, indiquant que la régulation de l'expression, un phénomène primordial dans l'établissement des phénotypes cellulaires, était en grande partie dirigée par la variabilité de la méthylation de l'ADN. Ces résultats sont importants puisqu'ils permettent d'appréhender le rôle potentiel de la méthylation de l'ADN, et de l'épigénétique en général, dans les phénomènes d'adaptation.

Par ailleurs, comme décrit par Etienne Danchin (Danchin, Pocheville, & Huneman, 2019), il existe un paradoxe étonnant vis à vis de l'hérédité des caractères acquis, une idée considérée comme une vision "Lamarckienne" et qui a été négligée au cours des dernières années au fur et à mesure du développement des théories dites "néo-Darwinniennes" de l'évolution. En effet, d'un point de vue purement Darwiniste, on peut considérer que les organismes devraient avoir évolué des moyens de transmettre les caractères qu'ils acquièrent

au cours de leur vie. Cette hypothèse vient du constat que nos ancêtres récents ont fait face à un environnement bien plus proche de celui que nous devrons affronter au cours de notre vie, en comparaison avec celui auquel ont été confrontés nos ancêtres lointains, distants de plusieurs milliers de générations, et dont nous héritons les séquences d'ADN. En conséquence, si un organisme venait à développer un moyen de transférer à sa descendance des traits acquis au cours de sa vie, via ses interactions avec son environnement, alors cet organisme serait favorisé par les processus de sélection naturelle vis à vis des organismes uniquement capables de transmettre leurs gènes (dans le sens de leurs séquences d'ADN). Ainsi, selon cette hypothèse, la transmission aux générations suivantes de nos différentes acclimatations à l'environnement actuel devraient probablement influencer l'expression des gènes en interaction avec notre environnement, à la manière des résultats que nous venons de décrire (Xu et al., 2015).

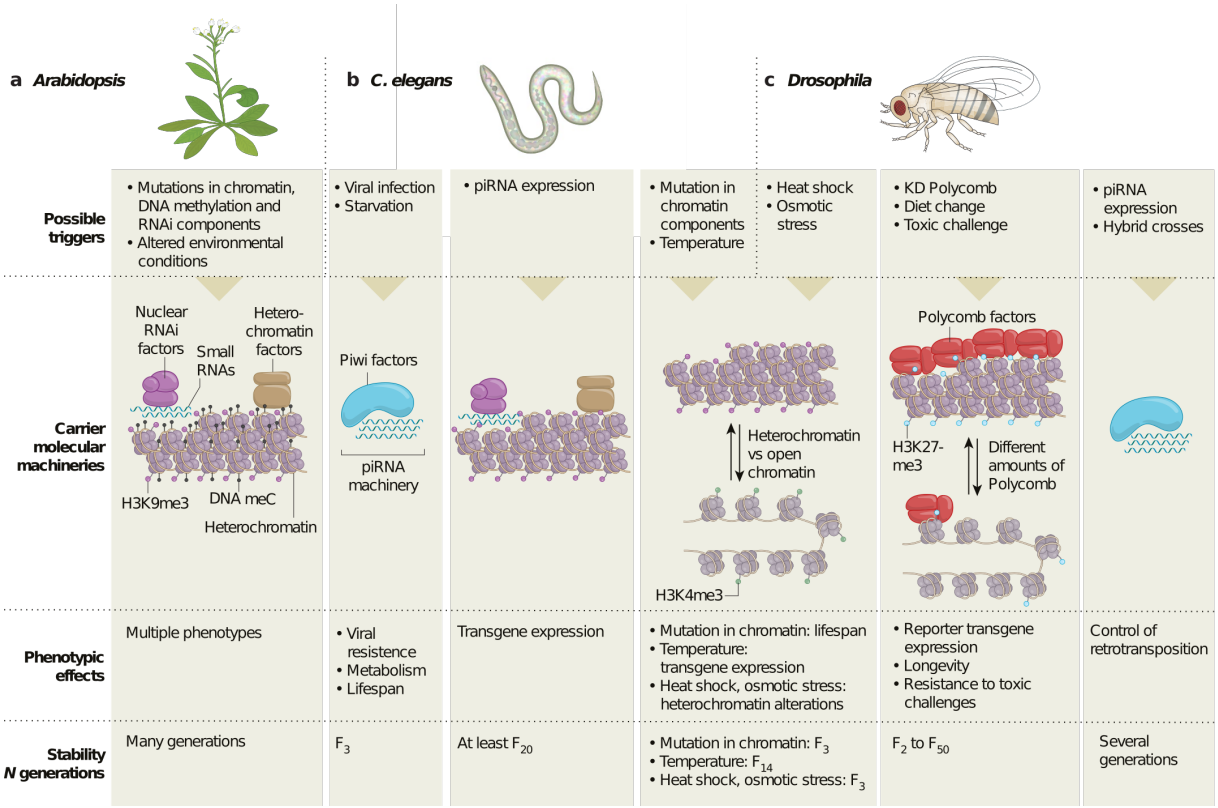

Fig. 7.1 Exemples d'hérédité épigénétique trans-générationnelle, chez différentes espèces

De haut en bas, la figure montre les mécanismes déclencheurs, les molécules impliquées dans la transmission de la mémoire trans-générationnelle, les conséquences phénotypiques des modifications épigénétiques et enfin le nombre de générations pendant lequel cette mémoire persiste, chez a) les plantes, b) C. elegans, et c) la drosophile. Figure tirée de (Cavalli & Heard, 2019).

Pris ensemble, les deux points que nous venons d'aborder nous permettent d'imaginer que c'est la variabilité épigénétique qui peut être responsable de l'établissement des phénotypes adaptatifs, au travers de son rôle régulateur de l'expression. Cette théorie, que nous avons abordée à la fin de la dernière partie, a précédemment été envisagée, en particulier par Klironomos et ses collègues en 2013 (Klironomos et al., 2013). Dans cette étude, les auteurs ont développé un modèle évolutif prenant en compte les effets combinés de la

génétique et de l'épigénétique en réponse à l'environnement. Ainsi, ils ont pu décrire comment la transmission d'information épigénétique pure pouvait affecter notre adaptation, au travers d'interactions entre variation génétique et épigénétique. Pour résumer les résultats de leur modèle, ils ont prédit différentes façons pour des mutations épigénétiques d'affecter les phénomènes d'adaptation. En premier lieu, leurs analyses suggèrent que lorsque les "variations héritables" correspondent à des variations épigénétiques, un découplage s'opère entre les variations génétiques et les variations de *fitness*. Ainsi, comme dans l'exemple de *Miscanthus lutarioriparius*, les phénotypes adaptatifs apparaissent avant les génotypes adaptatifs. Ensuite, la diversité génétique augmente progressivement au fur et à mesure qu'un phénotype adaptatif, codé par des variations épigénétiques, est mis en place. Enfin, dans cette situation, les mutations génétiques peuvent s'accumuler, comme sous neutralité, sans pression de l'environnement dans des populations de taille maintenue stable par l'effet des variations épigénétiques. Notons pour conclure que, s'il ne fait pas de doutes que les mécanismes d'évolution sur le long terme reposent sur une modification de la séquence d'ADN, ce modèle propose que des phénomènes d'hérédité non-génétique pourraient avoir un effet notable sur la façon dons nous nous adaptons (au sens évolutif) à notre environnement.

7.2.3 Le problème de l'héritabilité épigénétique transgénérationnelle

Dans la partie précédente nous avons discuté de la manière dont l'intégration de l'épigénétique aux modèles évolutifs pourrait nous permettre d'améliorer notre compréhension des phénomènes d'adaptation phénotypique. Toutefois, un pré-requis primordial aux résultats que nous venons de présenter semble être que les modifications épigénétiques puissent être transmises d'une génération à l'autre. Tout d'abord, dissocions deux formes d'hérédité épigénétique : en effet, si une modification de l'environnement peut affecter l'épigénome d'un individu adulte, alors l'épigénome de la lignée germinale de cet individu, ainsi que de celle du fœtus chez les femelles enceintes, peut être modifié aussi. En conséquence, on distinguera l'hérédité inter-générationnelle, regroupant les transmissions de l'épigénome à la génération F1 chez les mâles et F2 chez les femelles, de l'hérédité trans-générationnelle qui décrit l'ensemble des transmissions épigénétiques au delà de la première génération chez les mâles, et de la deuxième génération chez les femelles (Heard & Martienssen, 2014). S'il existe de nombreuses études décrivant des phénomènes d'hérédité inter-générationnelle chez les plantes en particulier mais aussi certaines espèces animales, les exemples d'hérédité trans-générationnelle sont bien plus rares (Heard & Martienssen, 2014 ; Jablonka, 2017 ; Cavalli & Heard, 2019) (figure 7.1). En outre, dans les quelques cas documentés d'hérédité trans-générationnelle, il est difficile d'écarter sans aucun doute l'influence de variation génétique sous-jacente, et le potentiel adaptatif de ces exemples laisse sceptiques une partie de la communauté scientifique (Heard & Martienssen, 2014 ; Horsthemke, 2018).

Toutefois, pour réconcilier les résultats des modèles évolutifs présentés dans la partie précédente avec cette barrière trans-générationnelle à laquelle font face les modifications épigénétiques, il est possible d'imaginer une sorte d'effet Baldwin se produisant à l'échelle des gènes, plus communément appelée "assimilation génétique". Dans cette hypothèse, sous l'effet d'une modification de l'environnement, des variations épigénétiques permettant une

acclimatation phénotypique peuvent émerger, et par hérédité inter-générationnelle, être transmises à la génération suivante. Si l'effet de l'environnement persiste, ce phénomène se répétera de génération en génération, jusqu'au retour à l'environnement initial, laissant ainsi potentiellement le temps pour l'apparition de mutations de la séquence de l'ADN "fixant" le phénotype adaptatif. Notons que l'effet de la mutation génétique sur le phénotype peut avoir lieu via la fixation des modifications épigénétiques initialement induites par l'environnement ou par un autre mécanisme indépendant. Ce modèle qui rejoint les conclusions présentées précédemment, permettrait aussi d'expliquer l'enrichissement considérable des eQTM en sites CpG sous contrôle génétique que nous avons détecté dans le chapitre 5. Enfin, pour renforcer cette théorie, puisque les marques épigénétiques peuvent contribuer à un taux de mutation plus élevé, il a été proposé que cette hérédité inter-générationnelle pourrait affecter la stabilité du génome et ainsi diriger l'apparition des mutations dans les régions d'intérêt (Danchin, Pocheville, & Huneman, 2019).

L'absence de preuves irréfutables en faveur de l'hérédité épigénétique transgénérationnelle est présentée par de nombreux défenseurs de la théorie actuelle de l'évolution comme un argument majeur à l'encontre de l'hérédité des caractères acquis et plus généralement d'une refonte de la théorie synthétique de l'évolution. Néanmoins, nous venons de proposer un modèle s'inscrivant dans les principes Darwinistes de sélection naturelle des variants génétiques, via une mutagénèse dirigée par l'activité des facteurs épigénétiques, et ce sans avoir recours à l'hérédité trans-générationnelle. En allant un pas plus loin, Etienne Danchin a proposé un modèle d'*Epigenetically-Facilitated Mutational Assimilation* dans lequel l'épigénétique joue le rôle d'une plaque tournante permettant l'assimilation génétique de toutes les formes d'hérédité non-génétique (Danchin, Pocheville, Rey, et al., 2019)(voir figure 7.2). En particulier, ce modèle suggère que l'ensemble des mécanismes d'hérédité non-génétique puissent agir à différentes échelles de temps, et ainsi former l'équivalent d'une course de relais au cours de laquelle le bâton transmis d'un partenaire à l'autre serait les informations de l'environnement permettant aux organismes de s'y adapter. L'épigénétique permettrait alors d'intégrer toutes les informations induites par les différents phénomènes d'hérédité non-génétique. En conséquence, à l'échelle de l'individu l'épigénétique serait responsable de l'adaptation aux conditions locales via son activité régulatrice de l'expression des gènes, tout en établissant un terrain fertile à l'apparition de mutations génétiques, et donc dirigerait à l'échelle inter-générationnelle les changements évolutifs de la séquence de l'ADN vers les régions fonctionnelles pertinentes aux conditions environnementales actuelles. Ce modèle plaide en faveur d'une intégration plus inclusive des mécanismes d'hérédité, qu'ils soient génétiques ou non-génétiques, à la théorie synthétique de l'évolution, une vision qui devrait permettre d'améliorer notre compréhension des mécanismes par lesquels les organismes s'adaptent aux modifications de leur environnement, l'un des enjeux majeurs de la biologie.

Pour conclure, je suis convaincu qu'à mesure que se développera notre compréhension des mécanismes de régulation de l'expression des gènes, et donc *in fine* de l'établissement des phénotypes, en particulier dans le contexte des différences entre populations humaines, de plus en plus d'arguments en faveur d'une nouvelle synthèse de la théorie de l'évolution émergeront. Une nouvelle synthèse qui permettrait peut-être de rendre justice aux idées associées à une vision Lamarckienne de l'évolution, comme sous-entendu par Bernhard Horsthemke : *Yes, Lamarck has never been dead and every so often raises his head, this time with the help of epigenetics* (Horsthemke, 2018).

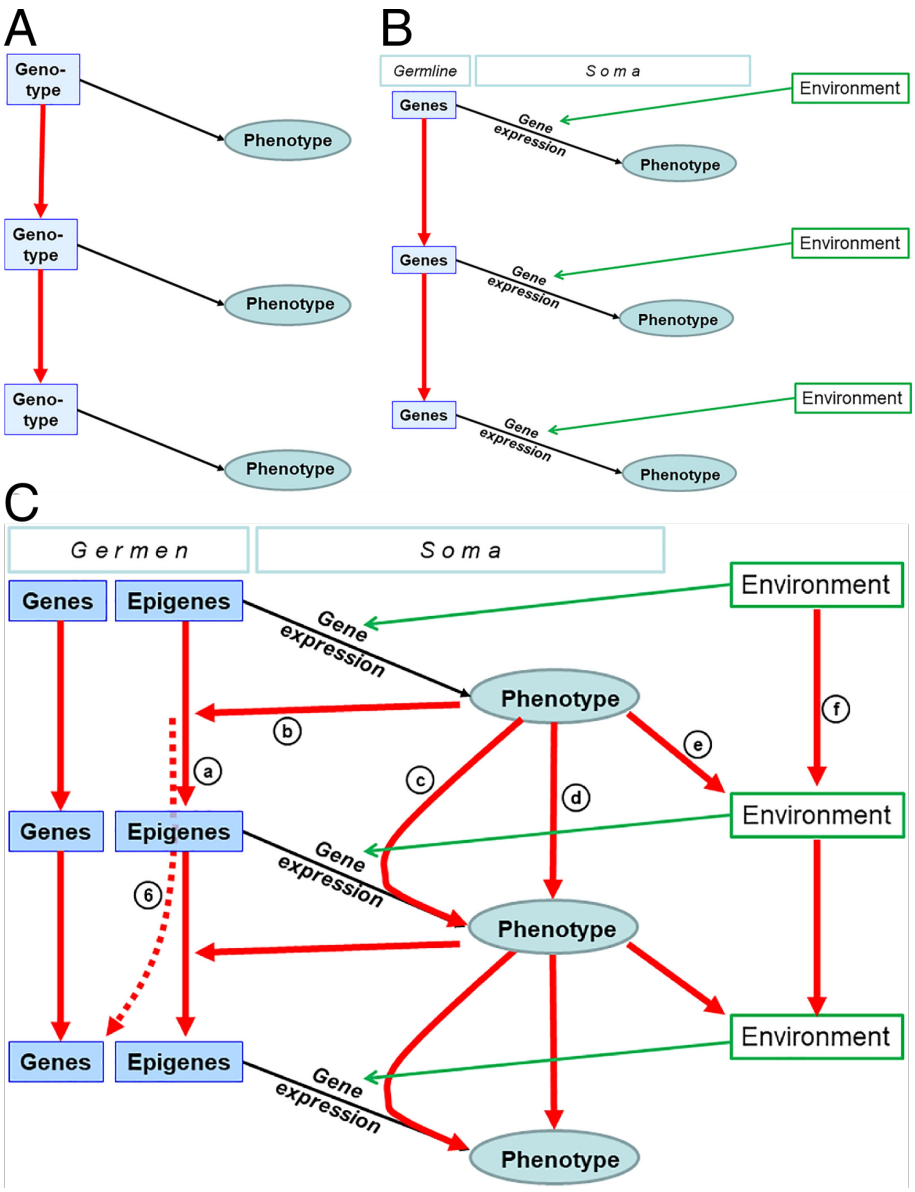

Fig. 7.2 Évolution des modèles descriptifs des processus d'hérédité
Les flèches noires correspondent au développement, les flèches rouges pleines à l'hérédité inter-générationnelle, les flèches rouges en pointillé à l'assimilation génétique qui se produit durant plusieurs générations, et les flèches vertes aux effets environnementaux. A) Vision de Maynard Smith (1965). B) Vision canonique de la théorie synthétique de l'évolution, où les gènes de la lignée germinale constituent la seule source d'information transmise d'une génération à l'autre. C) Vision émergente de l'hérédité inclusive décrivant : (6) epigenetically facilitated mutational assimilation ; (a) l'hérédité épigénétique de la lignée germinale ; (b) la communication soma-lignée germinale ; (c) les effets parentaux ; (d) l'hérédité culturelle ; (e) la construction de niches ; (f) l'hérédité écologique. Figure adaptée de (Danchin, Pocheville, Rey, et al., 2019).

Références

Aberg, K. A., McClay, J. L., Nerella, S., Clark, S., Kumar, G., Chen, W., Khachane, A. N., Xie, L., Hudson, A., Gao, G., Harada, A., Hultman, C. M., Sullivan, P. F., Magnusson, P. K., & van den Oord, E. J. (2014). Methylome-wide association study of schizophrenia : identifying blood biomarker signatures of environmental insults. *JAMA Psychiatry*, *71*(3), 255-64.

Abzhanov, A., Kuo, W. P., Hartmann, C., Grant, B. R., Grant, P. R., & Tabin, C. J. (2006). The calmodulin pathway and evolution of elongated beak morphology in darwin's finches. *Nature*, *442*(7102), 563-7.

Aguilera, O., Fernandez, A. F., Munoz, A., & Fraga, M. F. (2010). Epigenetics and environment : a complex relationship. *J Appl Physiol (1985)*, *109*(1), 243-51.

Alberts, B. (2002). *Molecular biology of the cell* (4th éd.). New York : Garland Science.

Annunziato, A. (2008). Dna packaging : Nucleosomes and chromatin. *Nature Education*, *1*(1), 26.

Aran, D., Toperoff, G., Rosenberg, M., & Hellman, A. (2011). Replication timing-related and gene body-specific methylation of active human genes. *Hum Mol Genet*, *20*(4), 670-80.

Arber, W. (1974). Dna modification and restriction. *Prog Nucleic Acid Res Mol Biol*, *14*(0), 1-37.

Arita, K., Ariyoshi, M., Tochio, H., Nakamura, Y., & Shirakawa, M. (2008). Recognition of hemi-methylated dna by the sra protein uhrf1 by a base-flipping mechanism. *Nature*, *455*(7214), 818-21.

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Res*, *21*(3), 381-95.

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., Roux, J., Pritchard, J. K., & Gilad, Y. (2014). Methylation qtls are associated with coordinated changes in transcription factor binding, histone modifications, and gene expression levels. *PLoS Genet*, *10*(9), e1004663.

Barrow, T. M., & Michels, K. B. (2014). Epigenetic epidemiology of cancer. *Biochem Biophys Res Commun*, *455*(1-2), 70-83.

Bartel, D. P. (2009). Micrornas : target recognition and regulatory functions. *Cell*, *136*(2), 215-33.

Bednar, J., Horowitz, R. A., Grigoryev, S. A., Carruthers, L. M., Hansen, J. C., Koster, A. J., & Woodcock, C. L. (1998). Nucleosomes, linker dna, and linker histone form a unique structural motif that directs the higher-order folding and compaction of chromatin. *Proc Natl Acad Sci U S A*, *95*(24), 14173-8.

Bell, J. T., Loomis, A. K., Butcher, L. M., Gao, F., Zhang, B., Hyde, C. L., Sun, J., Wu, H., Ward, K., Harris, J., Scollen, S., Davies, M. N., Schalkwyk, L. C., Mill, J., Mu, T. C., Williams, F. M., Li, N., Deloukas, P., Beck, S., McMahon, S. B., Wang, J.,

John, S. L., & Spector, T. D. (2014). Differential methylation of the trpa1 promoter in pain sensitivity. *Nat Commun*, *5*, 2978.

Bell, J. T., Pai, A. A., Pickrell, J. K., Gaffney, D. J., Pique-Regi, R., Degner, J. F., Gilad, Y., & Pritchard, J. K. (2011). Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome Biol*, *12*(1), R10.

Bell, J. T., & Spector, T. D. (2012). Dna methylation studies using twins : what are they telling us ? *Genome Biol*, *13*(10), 172.

Bellucci, E., Bitocchi, E., Ferrarini, A., Benazzo, A., Biagetti, E., Klie, S., Minio, A., Rau, D., Rodriguez, M., Panziera, A., Venturini, L., Attene, G., Albertini, E., Jackson, S. A., Nanni, L., Fernie, A. R., Nikoloski, Z., Bertorelle, G., Delledonne, M., & Papa, R. (2014). Decreased nucleotide and expression diversity and modified coexpression patterns characterize domestication in the common bean. *Plant Cell*, *26*(5), 1901-1912.

Benner, S. A. (2010). Defining life. *Astrobiology*, *10*(10), 1021-30.

Bhatlekar, S., Fields, J. Z., & Boman, B. M. (2014). Hox genes and their role in the development of human cancers. *J Mol Med (Berl)*, *92*(8), 811-23.

Bird, A. (2007). Perceptions of epigenetics. *Nature*, *447*(7143), 396-8.

Bird, A., Taggart, M., Frommer, M., Miller, O. J., & Macleod, D. (1985). A fraction of the mouse genome that is derived from islands of nonmethylated, cpg-rich dna. *Cell*, *40*(1), 91-9.

Bird, A. P. (1986). Cpg-rich islands and the function of dna methylation. *Nature*, *321*(6067), 209-13.

Bird, A. P. (1987). Cpg islands as gene markers in the vertebrate nucleus. *Trends in Genetics*, *3*, 342-347.

Birney, E., Smith, G. D., & Greally, J. M. (2016). Epigenome-wide association studies and the interpretation of disease -omics. *PLOS Genetics*, *12*(6), e1006105.

Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., Yu, W., Rongione, M. A., Ekstrom, T. J., Harris, T. B., Launer, L. J., Eiriksdottir, G., Leppert, M. F., Sapienza, C., Gudnason, V., & Feinberg, A. P. (2008). Intra-individual change over time in dna methylation with familial clustering. *JAMA*, *299*(24), 2877-83.

Blackwood, E. M., & Kadonaga, J. T. (1998). Going the distance : a current view of enhancer action. *Science*, *281*(5373), 60-3.

B Larsen, B., Miller, E., K Rhodes, M., & Wiens, J. (2017). *Inordinate fondness multiplied and redistributed : the number of species on earth and the new pie of life* (Vol. 92).

Blow, M. J., Clark, T. A., Daum, C. G., Deutschbauer, A. M., Fomenkov, A., Fries, R., Froula, J., Kang, D. D., Malmstrom, R. R., Morgan, R. D., Posfai, J., Singh, K., Visel, A., Wetmore, K., Zhao, Z., Rubin, E. M., Korlach, J., Pennacchio, L. A., & Roberts, R. J. (2016). The epigenomic landscape of prokaryotes. *PLoS Genet*, *12*(2), e1005854.

Bock, C., Walter, J., Paulsen, M., & Lengauer, T. (2008). Inter-individual variation of dna methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res*, *36*(10), e55.

Bollati, V., Schwartz, J., Wright, R., Litonjua, A., Tarantini, L., Suh, H., Sparrow, D., Vokonas, P., & Baccarelli, A. (2009). Decline in genomic dna methylation through aging in a cohort of elderly subjects. *Mech Ageing Dev*, *130*(4), 234-9.

Bonder, M. J., Kasela, S., Kals, M., Tamm, R., Lokk, K., Barragan, I., Buurman, W. A., Deelen, P., Greve, J. W., Ivanov, M., Rensen, S. S., van Vliet-Ostaptchouk, J. V.,

Wolfs, M. G., Fu, J., Hofker, M. H., Wijmenga, C., Zhernakova, A., Ingelman-Sundberg, M., Franke, L., & Milani, L. (2014). Genetic and epigenetic regulation of gene expression in fetal and adult human livers. *BMC Genomics*, *15*, 860.

Bonder, M. J., Luijk, R., Zhernakova, D. V., Moed, M., Deelen, P., Vermaat, M., van Iterson, M., van Dijk, F., van Galen, M., Bot, J., Slieker, R. C., Jhamai, P. M., Verbiest, M., Suchiman, H. E., Verkerk, M., van der Breggen, R., van Rooij, J., Lakenberg, N., Arindrarto, W., Kielbasa, S. M., Jonkers, I., van 't Hof, P., Nooren, I., Beekman, M., Deelen, J., van Heemst, D., Zhernakova, A., Tigchelaar, E. F., Swertz, M. A., Hofman, A., Uitterlinden, A. G., Pool, R., van Dongen, J., Hottenga, J. J., Stehouwer, C. D., van der Kallen, C. J., Schalkwijk, C. G., van den Berg, L. H., van Zwet, E. W., Mei, H., Li, Y., Lemire, M., Hudson, T. J., Consortium, B., Slagboom, P. E., Wijmenga, C., Veldink, J. H., van Greevenbroek, M. M., van Duijn, C. M., Boomsma, D. I., Isaacs, A., Jansen, R., van Meurs, J. B., t Hoen, P. A., Franke, L., & Heijmans, B. T. (2017). Disease variants alter transcription factor levels and methylation of their binding sites. *Nat Genet*, *49*(1), 131-138.

Bootsma, H. J., Egmont-Petersen, M., & Hermans, P. W. (2007). Analysis of the in vitro transcriptional response of human pharyngeal epithelial cells to adherent streptococcus pneumoniae : evidence for a distinct response to encapsulated strains. *Infect Immun*, *75*(11), 5489-99.

Bostick, M., Kim, J. K., Esteve, P. O., Clark, A., Pradhan, S., & Jacobsen, S. E. (2007). Uhrf1 plays a role in maintaining dna methylation in mammalian cells. *Science*, *317*(5845), 1760-4.

Bouchard-Mercier, A., Paradis, A. M., Rudkowska, I., Lemieux, S., Couture, P., & Vohl, M. C. (2013). Associations between dietary patterns and gene expression profiles of healthy men and women : a cross-sectional study. *Nutr J*, *12*, 24.

Boutens, L., & Stienstra, R. (2016). Adipose tissue macrophages : going off track during obesity. *Diabetologia*, *59*(5), 879-94.

Breschi, A., Djebali, S., Gillis, J., Pervouchine, D. D., Dobin, A., Davis, C. A., Gingeras, T. R., & Guigó, R. (2016). Gene-specific patterns of expression variation across organs and species. *Genome Biology*, *17*(1), 151.

Budhavarapu, V., Chavez, M., & K Tyler, J. (2013). *How is epigenetic information maintained through dna replication ?* (Vol. 6).

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, *144*(3), 327-39.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousgou, O., Whetzel, P. L., Amode, R., Guillen, J. A., Riat, H. S., Trevanion, S. J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorff, L. A., Cunningham, F., & Parkinson, H. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, *47*(D1), D1005-D1012.

Burkhardt, J., R. W. (2013). Lamarck, evolution, and the inheritance of acquired characters. *Genetics*, *194*(4), 793-805.

Cameron-Smith, D., Burke, L. M., Angus, D. J., Tunstall, R. J., Cox, G. R., Bonen, A., Hawley, J. A., & Hargreaves, M. (2003). A short-term, high-fat diet up-regulates lipid metabolism and gene expression in human skeletal muscle. *Am J Clin Nutr*, *77*(2), 313-8.

Cantone, I., & Fisher, A. G. (2013). Epigenetic programming and reprogramming during

development. *Nat Struct Mol Biol*, *20*(3), 282-9.

Cao, J. (2014). The functional role of long non-coding rnas and epigenetics. *Biol Proced Online*, *16*, 11.

Carcamo-Orive, I., Hoffman, G. E., Cundiff, P., Beckmann, N. D., D'Souza, S. L., Knowles, J. W., Patel, A., Papatsenko, D., Abbasi, F., Reaven, G. M., Whalen, S., Lee, P., Shahbazi, M., Henrion, M. Y. R., Zhu, K., Wang, S., Roussos, P., Schadt, E. E., Pandey, G., Chang, R., Quertermous, T., & Lemischka, I. (2017). Analysis of transcriptional variability in a large human ipsc library reveals genetic and non-genetic determinants of heterogeneity. *Cell Stem Cell*, *20*(4), 518-532 e9.

Cardoso-Moreira, M., Halbert, J., Valloton, D., Velten, B., Chen, C., Shao, Y., Liechti, A., Ascencao, K., Rummel, C., Ovchinnikova, S., Mazin, P. V., Xenarios, I., Harshman, K., Mort, M., Cooper, D. N., Sandi, C., Soares, M. J., Ferreira, P. G., Afonso, S., Carneiro, M., Turner, J. M. A., VandeBerg, J. L., Fallahshahroudi, A., Jensen, P., Behr, R., Lisgo, S., Lindsay, S., Khaitovich, P., Huber, W., Baker, J., Anders, S., Zhang, Y. E., & Kaessmann, H. (2019). Gene expression across mammalian organ development. *Nature*.

Carelli, F. N., Liechti, A., Halbert, J., Warnefors, M., & Kaessmann, H. (2018). Repurposing of promoters and enhancers during mammalian evolution. *Nat Commun*, *9*(1), 4066.

Carja, O., MacIsaac, J. L., Mah, S. M., Henn, B. M., Kobor, M. S., Feldman, M. W., & Fraser, H. B. (2017). Worldwide patterns of human epigenetic variation. *Nat Ecol Evol*, *1*(10), 1577-1583.

Casadesus, J., & Low, D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiol Mol Biol Rev*, *70*(3), 830-56.

Castro-Vargas, C., Linares-Lopez, C., Lopez-Torres, A., Wrobel, K., Torres-Guzman, J. C., Hernandez, G. A., Wrobel, K., Lanz-Mendoza, H., & Contreras-Garduno, J. (2017). Methylation on rna : A potential mechanism related to immune priming within but not across generations. *Front Microbiol*, *8*, 473.

Catalanotto, C., Cogoni, C., & Zardo, G. (2016). Microrna in control of gene expression : An overview of nuclear functions. *Int J Mol Sci*, *17*(10).

Cavalli, G., & Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature*, *571*(7766), 489-499.

Chalancon, G., Ravarani, C. N., Balaji, S., Martinez-Arias, A., Aravind, L., Jothi, R., & Babu, M. M. (2012). Interplay between gene expression noise and regulatory network architecture. *Trends Genet*, *28*(5), 221-32.

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, J., G., Shapiro, M. D., Brady, S. D., Southwick, A. M., Absher, D. M., Grimwood, J., Schmutz, J., Myers, R. M., Petrov, D., Jonsson, B., Schluter, D., Bell, M. A., & Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *Science*, *327*(5963), 302-5.

Chappell, C., Beard, C., Altman, J., Jaenisch, R., & Jacob, J. (2006). Dna methylation by dna methyltransferase 1 is critical for effector cd8 t cell expansion. *J Immunol*, *176*(8), 4562-72.

Charlesworth, J. C., Curran, J. E., Johnson, M. P., Goring, H. H., Dyer, T. D., Diego, V. P., Kent, J., J. W., Mahaney, M. C., Almasy, L., MacCluer, J. W., Moses, E. K., & Blangero, J. (2010). Transcriptomic epidemiology of smoking : the effect of smoking on gene expression in lymphocytes. *BMC Med Genomics*, *3*, 29.

Chen, Q., Yan, W., & Duan, E. (2016). Epigenetic inheritance of acquired traits through

sperm rnas and sperm rna modifications. *Nat Rev Genet*, *17*(12), 733-743.

Chen, Z., Li, S., Subramaniam, S., Shyy, J. Y., & Chien, S. (2017). Epigenetic regulation : A new frontier for biomedical engineers. *Annu Rev Biomed Eng*, *19*, 195-219.

Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M., & Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat Genet*, *33*(3), 422-5.

Christensen, B. C., Houseman, E. A., Marsit, C. J., Zheng, S., Wrensch, M. R., Wiemels, J. L., Nelson, H. H., Karagas, M. R., Padbury, J. F., Bueno, R., Sugarbaker, D. J., Yeh, R. F., Wiencke, J. K., & Kelsey, K. T. (2009). Aging and environmental exposures alter tissue-specific dna methylation dependent upon cpg island context. *PLoS Genet*, *5*(8), e1000602.

Chu, J. H., Hart, J. E., Chhabra, D., Garshick, E., Raby, B. A., & Laden, F. (2016). Gene expression network analyses in response to air pollution exposures in the trucking industry. *Environ Health*, *15*(1), 101.

Chuang, J. C., & Jones, P. A. (2007). Epigenetics and microrna. *Pediatr Res*, *61*(5 Pt 2), 24R-29R.

Chuang, L. S., Ian, H. I., Koh, T. W., Ng, H. H., Xu, G., & Li, B. F. (1997). Human dna-(cytosine-5) methyltransferase-pcna complex as a target for p21waf1. *Science*, *277*(5334), 1996-2000.

Chung, H., Loehlin, D. W., Dufour, H. D., Vaccarro, K., Millar, J. G., & Carroll, S. B. (2014). A single gene affects both ecological divergence and mate choice in drosophila. *Science*, *343*(6175), 1148.

Coetzee, S. G., Pierce, S., Brundin, P., Brundin, L., Hazelett, D. J., & Coetzee, G. A. (2016). Enrichment of risk snps in regulatory regions implicate diverse tissues in parkinson's disease etiology. *Scientific Reports*, *6*, 30509.

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szczesniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for rna-seq data analysis. *Genome Biol*, *17*, 13.

Consortium, G. T. (2015). Human genomics. the genotype-tissue expression (gtex) pilot analysis : multitissue gene regulation in humans. *Science*, *348*(6235), 648-60.

Crick, F. H., Barnett, L., Brenner, S., & Watts-Tobin, R. J. (1961). General nature of the genetic code for proteins. *Nature*, *192*, 1227-32.

Da Costa, E. M., McInnes, G., Beaudry, A., & Raynal, N. J. (2017). Dna methylation-targeted drugs. *Cancer J*, *23*(5), 270-276.

Danchin, E., Pocheville, A., & Huneman, P. (2019). Early in life effects and heredity : reconciling neo-darwinism with neo-lamarckism under the banner of the inclusive evolutionary synthesis. *Philosophical Transactions of the Royal Society B : Biological Sciences*, *374*(1770), 20180113.

Danchin, E., Pocheville, A., Rey, O., Pujol, B., & Blanchet, S. (2019). Epigenetically facilitated mutational assimilation : epigenetics as a hub within the inclusive evolutionary synthesis. *Biological Reviews*, *94*(1), 259-282.

Darwin, C. (1859). *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life.* London : John Murray.

Davey Smith, G., & Hemani, G. (2014). Mendelian randomization : genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*, *23*(R1), R89-98.

Dawkins, R. (1976). *The selfish gene.* Oxford : Oxford University Press.

Daxinger, L., & Whitelaw, E. (2012). Understanding transgenerational epigenetic inheritance via the gametes in mammals. *Nat Rev Genet*, *13*(3), 153-62.

Dean, A. (2006). On a chromosome far, far away : Lcrs and gene expression. *Trends Genet*, *22*(1), 38-45.

Deaton, A. M., & Bird, A. (2011). Cpg islands and the regulation of transcription. *Genes Dev*, *25*(10), 1010-22.

Delaneau, O., Zazhytska, M., Borel, C., Giannuzzi, G., Rey, G., Howald, C., Kumar, S., Ongen, H., Popadin, K., Marbach, D., Ambrosini, G., Bielser, D., Hacker, D., Romano, L., Ribaux, P., Wiederkehr, M., Falconnet, E., Bucher, P., Bergmann, S., Antonarakis, S. E., Reymond, A., & Dermitzakis, E. T. (2019). Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science*, *364*(6439), eaat8266.

Dion-Cote, A. M., Renaut, S., Normandeau, E., & Bernatchez, L. (2014). Rna-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Mol Biol Evol*, *31*(5), 1188-99.

Du, Q., Luu, P. L., Stirzaker, C., & Clark, S. J. (2015). Methyl-cpg-binding domain proteins : readers of the epigenome. *Epigenomics*, *7*(6), 1051-73.

Esteller, M. (2011). Non-coding rnas in human disease. *Nat Rev Genet*, *12*(12), 861-74.

Fagny, M., Patin, E., MacIsaac, J. L., Rotival, M., Flutre, T., Jones, M. J., Siddle, K. J., Quach, H., Harmant, C., McEwen, L. M., Froment, A., Heyer, E., Gessain, A., Betsem, E., Mouguiama-Daouda, P., Hombert, J.-M., Perry, G. H., Barreiro, L. B., Kobor, M. S., & Quintana-Murci, L. (2015). The epigenomic landscape of african rainforest hunter-gatherers and farmers. *Nature Communications*, *6*, 10047.

Fave, M. J., Lamaze, F. C., Soave, D., Hodgkinson, A., Gauvin, H., Bruat, V., Grenier, J. C., Gbeha, E., Skead, K., Smargiassi, A., Johnson, M., Idaghdour, Y., & Awadalla, P. (2018). Gene-by-environment interactions in urban populations modulate risk phenotypes. *Nat Commun*, *9*(1), 827.

Feng, S., Jacobsen, S. E., & Reik, W. (2010). Epigenetic reprogramming in plant and animal development. *Science*, *330*(6004), 622-7.

Fraser, H. B. (2013). Gene expression drives local adaptation in humans. *Genome Res*, *23*(7), 1089-96.

Fraser, H. B., Lam, L. L., Neumann, S. M., & Kobor, M. S. (2012). Population-specificity of human dna methylation. *Genome Biol*, *13*(2), R8.

Garg, P., Joshi, R. S., Watson, C., & Sharp, A. J. (2018). A survey of inter-individual variation in dna methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLoS Genet*, *14*(10), e1007707.

Gassmann, R., Vagnarelli, P., Hudson, D., & Earnshaw, W. C. (2004). Mitotic chromosome formation and the condensin paradox. *Exp Cell Res*, *296*(1), 35-42.

Genomes Project, C., Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E., & McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061-73.

Gerstein, M. B., Rozowsky, J., Yan, K. K., Wang, D., Cheng, C., Brown, J. B., Davis, C. A., Hillier, L., Sisu, C., Li, J. J., Pei, B., Harmanci, A. O., Duff, M. O., Djebali, S., Alexander, R. P., Alver, B. H., Auerbach, R., Bell, K., Bickel, P. J., Boeck, M. E., Boley, N. P., Booth, B. W., Cherbas, L., Cherbas, P., Di, C., Dobin, A., Drenkow, J., Ewing, B., Fang, G., Fastuca, M., Feingold, E. A., Frankish, A., Gao, G., Good, P. J., Guigo, R., Hammonds, A., Harrow, J., Hoskins, R. A., Howald, C., Hu, L., Huang, H., Hubbard, T. J., Huynh, C., Jha, S., Kasper, D., Kato, M., Kaufman, T. C., Kitchen, R. R., Ladewig, E., Lagarde, J., Lai, E., Leng, J., Lu, Z., MacCoss, M., May, G., McWhirter, R., Merrihew, G., Miller, D. M., Mortazavi, A., Murad,

R., Oliver, B., Olson, S., Park, P. J., Pazin, M. J., Perrimon, N., Pervouchine, D., Reinke, V., Reymond, A., Robinson, G., Samsonova, A., Saunders, G. I., Schlesinger, F., Sethi, A., Slack, F. J., Spencer, W. C., Stoiber, M. H., Strasbourger, P., Tanzer, A., Thompson, O. A., Wan, K. H., Wang, G., Wang, H., Watkins, K. L., Wen, J., Wen, K., Xue, C., Yang, L., Yip, K., Zaleski, C., Zhang, Y., Zheng, H., Brenner, S. E., Graveley, B. R., Celniker, S. E., Gingeras, T. R., & Waterston, R. (2014). Comparative analysis of the transcriptome across distant species. *Nature*, *512*(7515), 445-8.

Gluckman, P. D., Hanson, M. A., Buklijas, T., Low, F. M., & Beedle, A. S. (2009). Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *Nat Rev Endocrinol*, *5*(7), 401-8.

Grandbois, M., Beyer, M., Rief, M., Clausen-Schaumann, H., & Gaub, H. E. (1999). How strong is a covalent bond? *Science*, *283*(5408), 1727-30.

Grath, S., & Parsch, J. (2016). Sex-biased gene expression. *Annual Review of Genetics*, *50*(1), 29-44.

Grigoryev, S. A., & Woodcock, C. L. (2012). Chromatin organization - the 30 nm fiber. *Exp Cell Res*, *318*(12), 1448-55.

Gronniger, E., Weber, B., Heil, O., Peters, N., Stab, F., Wenck, H., Korn, B., Winnefeld, M., & Lyko, F. (2010). Aging and chronic sun exposure cause distinct epigenetic changes in human skin. *PLoS Genet*, *6*(5), e1000971.

Grundberg, E., Meduri, E., Sandling, J. K., Hedman, A. K., Keildson, S., Buil, A., Busche, S., Yuan, W., Nisbet, J., Sekowska, M., Wilk, A., Barrett, A., Small, K. S., Ge, B., Caron, M., Shin, S. Y., Multiple Tissue Human Expression Resource, C., Lathrop, M., Dermitzakis, E. T., McCarthy, M. I., Spector, T. D., Bell, J. T., & Deloukas, P. (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*(5), 876-90.

Guo, J. U., Ma, D. K., Mo, H., Ball, M. P., Jang, M. H., Bonaguidi, M. A., Balazer, J. A., Eaves, H. L., Xie, B., Ford, E., Zhang, K., Ming, G. L., Gao, Y., & Song, H. (2011). Neuronal activity modifies the dna methylation landscape in the adult brain. *Nat Neurosci*, *14*(10), 1345-51.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S. B., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S. E., & Dermitzakis, E. T. (2013). Passive and active dna methylation and the interplay with genetic variation in gene regulation. *Elife*, *2*, e00523.

Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S. B., Buil, A., Yurovsky, A., Bryois, J., Padioleau, I., Romano, L., Planchon, A., Falconnet, E., Bielser, D., Gagnebin, M., Giger, T., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S. E., & Dermitzakis, E. T. (2015). Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, *11*(1), e1004958.

Haerty, W., & Singh, R. S. (2006). Gene regulation divergence is a major contributor to the evolution of dobzhansky-muller incompatibilities between species of drosophila. *Mol Biol Evol*, *23*(9), 1707-14.

Hannon, E., Knox, O., Sugden, K., Burrage, J., Wong, C. C. Y., Belsky, D. W., Corcoran, D. L., Arseneault, L., Moffitt, T. E., Caspi, A., & Mill, J. (2018). Characterizing

genetic and environmental influences on variable dna methylation using monozygotic and dizygotic twins. *PLoS Genet*, *14*(8), e1007544.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J. B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., & Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*, *49*(2), 359-367.

Hansen, K. H., Bracken, A. P., Pasini, D., Dietrich, N., Gehani, S. S., Monrad, A., Rappsilber, J., Lerdrup, M., & Helin, K. (2008). A model for transmission of the h3k27me3 epigenetic mark. *Nat Cell Biol*, *10*(11), 1291-300.

Harris, M. B., Mostecki, J., & Rothman, P. B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative bcl-6 function. *J Biol Chem*, *280*(13), 13114-21.

Harris, S. E., Riggio, V., Evenden, L., Gilchrist, T., McCafferty, S., Murphy, L., Wrobel, N., Taylor, A. M., Corley, J., Pattie, A., Cox, S. R., Martin-Ruiz, C., Prendergast, J., Starr, J. M., Marioni, R. E., & Deary, I. J. (2017). Age-related gene expression changes, and transcriptome wide association study of physical and cognitive aging traits, in the lothian birth cohort 1936. *Aging (Albany NY)*, *9*(12), 2489-2503.

Harrison, G. F., Sanz, J., Boulais, J., Mina, M. J., Grenier, J.-C., Leng, Y., Dumaine, A., Yotova, V., Bergey, C. M., Nsobya, S. L., Elledge, S. J., Schurr, E., Quintana-Murci, L., Perry, G. H., & Barreiro, L. B. (2019). Natural selection contributed to immunological differences between hunter-gatherers and agriculturalists. *Nature Ecology & Evolution*, *3*(8), 1253-1264.

Hashimoto, H., Liu, Y., Upadhyay, A. K., Chang, Y., Howerton, S. B., Vertino, P. M., Zhang, X., & Cheng, X. (2012). Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res*, *40*(11), 4841-9.

He, Y. F., Li, B. Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., Sun, Y., Li, X., Dai, Q., Song, C. X., Zhang, K., He, C., & Xu, G. L. (2011). Tet-mediated formation of 5-carboxylcytosine and its excision by tdg in mammalian dna. *Science*, *333*(6047), 1303-7.

Heard, E., & Martienssen, R. A. (2014). Transgenerational epigenetic inheritance : myths and mechanisms. *Cell*, *157*(1), 95-109.

Hermann, A., Goyal, R., & Jeltsch, A. (2004). The dnmt1 dna-(cytosine-c5)-methyltransferase methylates dna processively with high preference for hemimethylated target sites. *J Biol Chem*, *279*(46), 48350-9.

Herold, M., Bartkuhn, M., & Renkawitz, R. (2012). Ctcf : insights into insulator function during development. *Development*, *139*(6), 1045-57.

Heyn, H., Moran, S., Hernando-Herraez, I., Sayols, S., Gomez, A., Sandoval, J., Monk, D., Hata, K., Marques-Bonet, T., Wang, L., & Esteller, M. (2013). Dna methylation contributes to natural human variation. *Genome Res*, *23*(9), 1363-72.

Hirth, F., Hartmann, B., & Reichert, H. (1998). Homeotic gene action in embryonic brain development of drosophila. *Development*, *125*(9), 1579-89.

Holland, P. W., Booth, H. A., & Bruford, E. A. (2007). Classification and nomenclature of all human homeobox genes. *BMC Biol*, *5*, 47.

Holliday, R. (1990). Mechanisms for the control of gene activity during development. *Biol Rev Camb Philos Soc*, *65*(4), 431-71.

Horsthemke, B. (2018). A critical view on transgenerational epigenetic inheritance in humans. *Nat Commun*, *9*(1), 2973.

Horvath, S. (2013). Dna methylation age of human tissues and cell types. *Genome Biol*, *14*(10), R115.

Horvath, S., Erhart, W., Brosch, M., Ammerpohl, O., von Schonfels, W., Ahrens, M., Heits, N., Bell, J. T., Tsai, P. C., Spector, T. D., Deloukas, P., Siebert, R., Sipos, B., Becker, T., Rocken, C., Schafmayer, C., & Hampe, J. (2014). Obesity accelerates epigenetic aging of human liver. *Proc Natl Acad Sci U S A*, *111*(43), 15538-43.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., & Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*, *13*, 86.

Huang, S. (2009). Non-genetic heterogeneity of cells in development : more than just noise. *Development*, *136*(23), 3853-62.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhajarvi, T., Chia, J. M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeppler, S. M., Lai, J., Morrell, P. L., Shannon, L. M., Song, C., Springer, N. M., Swanson-Wagner, R. A., Tiffin, P., Wang, J., Zhang, G., Doebley, J., McMullen, M. D., Ware, D., Buckler, E. S., Yang, S., & Ross-Ibarra, J. (2012). Comparative population genomics of maize domestication and improvement. *Nat Genet*, *44*(7), 808-11.

Husquin, L. T., Rotival, M., Fagny, M., Quach, H., Zidane, N., McEwen, L. M., MacIsaac, J. L., Kobor, M. S., Aschard, H., Patin, E., & Quintana-Murci, L. (2018). Exploring the genetic basis of human population differences in dna methylation and their causal impact on immune gene regulation. *Genome Biol*, *19*(1), 222.

Irie, N., Satoh, N., & Kuratani, S. (2018). The phylum vertebrata : a case for zoological recognition. *Zoological Lett*, *4*, 32.

Issa, J. P. (2007). Dna methylation as a therapeutic target in cancer. *Clin Cancer Res*, *13*(6), 1634-7.

Jablonka, E. (2017). The evolutionary implications of epigenetic inheritance. *Interface Focus*, *7*(5), 20160135.

Jablonka, E., & Raz, G. (2009). Transgenerational epigenetic inheritance : prevalence, mechanisms, and implications for the study of heredity and evolution. *Q Rev Biol*, *84*(2), 131-76.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression : how the genome integrates intrinsic and environmental signals. *Nat Genet*, *33 Suppl*, 245-54.

J. Bowler, P. (1976). *Malthus, darwin, and the concept of struggle* (Vol. 37).

Jenkin, F. (1867). Review of the origin of species. *The North British Review*, *46*(June), 277-318.

Jin, B., Li, Y., & Robertson, K. D. (2011). Dna methylation : superior or subordinate in the epigenetic hierarchy ? *Genes Cancer*, *2*(6), 607-17.

Jjingo, D., Conley, A. B., Yi, S. V., Lunyak, V. V., & Jordan, I. K. (2012). On the presence and role of human gene-body dna methylation. *Oncotarget*, *3*(4), 462-74.

Jodar, M., Selvaraju, S., Sendler, E., Diamond, M. P., Krawetz, S. A., & Reproductive Medicine, N. (2013). The presence, role and clinical use of spermatozoal rnas. *Hum Reprod Update*, *19*(6), 604-24.

Jones, P. A. (2012). Functions of dna methylation : islands, start sites, gene bodies and beyond. *Nat Rev Genet*, *13*(7), 484-92.

Jones, P. A., Issa, J. P., & Baylin, S. (2016). Targeting the cancer epigenome for therapy. *Nat Rev Genet*, *17*(10), 630-41.

Juliano, C., Wang, J., & Lin, H. (2011). Uniting germline and stem cells : the function of piwi proteins and the pirna pathway in diverse organisms. *Annu Rev Genet*, *45*,

447-69.

Kader, F., & Ghai, M. (2017). Dna methylation-based variation between human populations. *Mol Genet Genomics*, *292*(1), 5-35.

Kalantry, S. (2011). Recent advances in x-chromosome inactivation. *J Cell Physiol*, *226*(7), 1714-8.

Kaplow, I. M., MacIsaac, J. L., Mah, S. M., McEwen, L. M., Kobor, M. S., & Fraser, H. B. (2015). A pooling-based approach to mapping genetic variants associated with dna methylation. *Genome Res*, *25*(6), 907-17.

Karaman, M. W., Houck, M. L., Chemnick, L. G., Nagpal, S., Chawannakul, D., Sudano, D., Pike, B. L., Ho, V. V., Ryder, O. A., & Hacia, J. G. (2003). Comparative analysis of gene-expression patterns in human and african great ape cultured fibroblasts. *Genome Res*, *13*(7), 1619-30.

Karlen, S. J., & Krubitzer, L. (2006). Phenotypic diversity is the cornerstone of evolution : variation in cortical field size within short-tailed opossums. *J Comp Neurol*, *499*(6), 990-9.

Kaufmann, E., Sanz, J., Dunn, J. L., Khan, N., Mendonca, L. E., Pacis, A., Tzelepis, F., Pernet, E., Dumaine, A., Grenier, J. C., Mailhot-Leonard, F., Ahmed, E., Belle, J., Besla, R., Mazer, B., King, I. L., Nijnik, A., Robbins, C. S., Barreiro, L. B., & Divangahi, M. (2018). Bcg educates hematopoietic stem cells to generate protective innate immunity against tuberculosis. *Cell*, *172*(1-2), 176-190 e19.

Kim, S., Yu, N. K., & Kaang, B. K. (2015). Ctcf as a multifunctional protein in genome regulation and gene expression. *Exp Mol Med*, *47*, e166.

Kim-Hellmuth, S., Bechheim, M., Putz, B., Mohammadi, P., Nedelec, Y., Giangreco, N., Becker, J., Kaiser, V., Fricker, N., Beier, E., Boor, P., Castel, S. E., Nothen, M. M., Barreiro, L. B., Pickrell, J. K., Muller-Myhsok, B., Lappalainen, T., Schumacher, J., & Hornung, V. (2017). Genetic regulatory effects modified by immune activation contribute to autoimmune disease associations. *Nat Commun*, *8*(1), 266.

King, M. C., & Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, *188*(4184), 107-16.

Klironomos, F. D., Berg, J., & Collins, S. (2013). How epigenetic mutations can affect genetic evolution : model and mechanism. *Bioessays*, *35*(6), 571-8.

Kmita, M., & Duboule, D. (2003). Organizing axes in time and space ; 25 years of colinear tinkering. *Science*, *301*(5631), 331-3.

Koenig, D., Jimenez-Gomez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., Kumar, R., Covington, M. F., Devisetty, U. K., Tat, A. V., Tohge, T., Bolger, A., Schneeberger, K., Ossowski, S., Lanz, C., Xiong, G., Taylor-Teeples, M., Brady, S. M., Pauly, M., Weigel, D., Usadel, B., Fernie, A. R., Peng, J., Sinha, N. R., & Maloof, J. N. (2013). Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc Natl Acad Sci U S A*, *110*(28), E2655-62.

Koestler, D. C., Christensen, B., Karagas, M. R., Marsit, C. J., Langevin, S. M., Kelsey, K. T., Wiencke, J. K., & Houseman, E. A. (2013). Blood-based profiles of dna methylation predict the underlying distribution of cell types : a validation analysis. *Epigenetics*, *8*(8), 816-26.

Kohli, R. M., & Zhang, Y. (2013). Tet enzymes, tdg and the dynamics of dna demethylation. *Nature*, *502*(7472), 472-9.

Komori, H. K., Hart, T., LaMere, S. A., Chew, P. V., & Salomon, D. R. (2015). Defining cd4 t cell memory by the epigenetic landscape of cpg dna methylation. *J Immunol*, *194*(4), 1565-79.

Korlach, J., & Turner, S. W. (2012). Going beyond five bases in dna sequencing. *Curr Opin Struct Biol*, *22*(3), 251-61.

Kornberg, T. B., & Tabata, T. (1993). Segmentation of the drosophila embryo. *Curr Opin Genet Dev*, *3*(4), 585-94.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell*, *128*(4), 693-705.

Kradolfer, D., Wolff, P., Jiang, H., Siretskiy, A., & Kohler, C. (2013). An imprinted gene underlies postzygotic reproductive isolation in arabidopsis thaliana. *Dev Cell*, *26*(5), 525-35.

Kulis, M., & Esteller, M. (2010). Dna methylation and cancer. *Adv Genet*, *70*, 27-56.

LaFlamme, B. (2014). Gene expression in early development. *Nature Genetics*, *46*, 99.

Lam, L. L., Emberly, E., Fraser, H. B., Neumann, S. M., Chen, E., Miller, G. E., & Kobor, M. S. (2012). Factors underlying variable dna methylation in a human community cohort. *Proc Natl Acad Sci U S A*, *109 Suppl 2*, 17253-60.

Lamarck, J.-B.-P.-A. (1809). *Philosophie zoologique* (Dentu éd.). Paris : Dentu, Libraire, 3 rue du Pont de Lodi.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860-921.

Lappalainen, T., & Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet*, *18*(7), 441-451.

Lappin, T. R., Grier, D. G., Thompson, A., & Halliday, H. L. (2006). Hox genes : seductive science, mysterious mechanisms. *Ulster Med J*, *75*(1), 23-31.

Lee, C. (2018). Genome-wide expression quantitative trait loci analysis using mixed models. *Frontiers in Genetics*, *9*, 341.

Lee, H. J., Hore, T. A., & Reik, W. (2014). Reprogramming the methylome : erasing memory and creating diversity. *Cell Stem Cell*, *14*(6), 710-9.

Lee, K. W., & Pausova, Z. (2013). Cigarette smoking and dna methylation. *Front Genet*, *4*, 132.

Lee, M. N., Ye, C., Villani, A. C., Raj, T., Li, W., Eisenhaure, T. M., Imboywa, S. H., Chipendo, P. I., Ran, F. A., Slowikowski, K., Ward, L. D., Raddassi, K., McCabe, C., Lee, M. H., Frohlich, I. Y., Hafler, D. A., Kellis, M., Raychaudhuri, S., Zhang, F., Stranger, B. E., Benoist, C. O., De Jager, P. L., Regev, A., & Hacohen, N. (2014).

Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science*, *343*(6175), 1246980.

Lee, P. P., Fitzpatrick, D. R., Beard, C., Jessup, H. K., Lehar, S., Makar, K. W., Perez-Melgosa, M., Sweetser, M. T., Schlissel, M. S., Nguyen, S., Cherry, S. R., Tsai, J. H., Tucker, S. M., Weaver, W. M., Kelso, A., Jaenisch, R., & Wilson, C. B. (2001). A critical role for dnmt1 and dna methylation in t cell development, function, and survival. *Immunity*, *15*(5), 763-74.

Lee, T. F., Zhai, J., & Meyers, B. C. (2010). Conservation and divergence in eukaryotic dna methylation. *Proc Natl Acad Sci U S A*, *107*(20), 9027-8.

Leroy, C.-G. (1802). *Lettres philosophiques sur l'intelligence et la perfectibilité des animaux, avec quelques lettres sur l'homme, par charles-georges leroy,... nouvelle édition, à laquelle on a joint des lettres posthumes sur l'homme, du même auteur. [avis de l'éditeur signé : Roux-fazillac.]*. Paris : Bossange Masson et Besson.

Levine, M., Cattoglio, C., & Tjian, R. (2014). Looping back to leap forward : transcription enters a new era. *Cell*, *157*(1), 13-25.

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., Hou, L., Baccarelli, A. A., Stewart, J. D., Li, Y., Whitsel, E. A., Wilson, J. G., Reiner, A. P., Aviv, A., Lohman, K., Liu, Y., Ferrucci, L., & Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, *10*(4), 573-591.

Lewis, E. B. (1978). A gene complex controlling segmentation in drosophila. *Nature*, *276*(5688), 565-70.

Li, E., Beard, C., & Jaenisch, R. (1993). Role for dna methylation in genomic imprinting. *Nature*, *366*(6453), 362-5.

Li, J., Liu, Y., Kim, T., Min, R., & Zhang, Z. (2010). Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Comput Biol*, *6*(8).

Li, L., He, S., Sun, J. M., & Davie, J. R. (2004). Gene regulation by sp1 and sp3. *Biochem Cell Biol*, *82*(4), 460-71.

Lim, C. Y., Santoso, B., Boulay, T., Dong, E., Ohler, U., & Kadonaga, J. T. (2004). The mte, a new core promoter element for transcription by rna polymerase ii. *Genes Dev*, *18*(13), 1606-17.

Lim, D. H. K., & Maher, E. R. (2010). Dna methylation : a form of epigenetic control of gene expression. *The Obstetrician & Gynaecologist*, *12*(1), 37-42.

Lister, R., Pelizzola, M., Dowen, R. H., Hawkins, R. D., Hon, G., Tonti-Filippini, J., Nery, J. R., Lee, L., Ye, Z., Ngo, Q.-M., Edsall, L., Antosiewicz-Bourget, J., Stewart, R., Ruotti, V., Millar, A. H., Thomson, J. A., Ren, B., & Ecker, J. R. (2009). Human dna methylomes at base resolution show widespread epigenomic differences. *Nature*, *462*, 315.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., Shchetynsky, K., Scheynius, A., Kere, J., Alfredsson, L., Klareskog, L., Ekstrom, T. J., & Feinberg, A. P. (2013). Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, *31*(2), 142-7.

Lloyd, E. (2001). *Units and levels of selection : An anatomy of the units of selection debates*.

Lobner-Olesen, A., Skovgaard, O., & Marinus, M. G. (2005). Dam methylation : coordinating cellular processes. *Curr Opin Microbiol*, *8*(2), 154-60.

Loeb, L. A., Loeb, K. R., & Anderson, J. P. (2003). Multiple mutations and cancer. *Proc*

Natl Acad Sci U S A, *100*(3), 776-81.

Lonard, D. M., & O'Malley, B. W. (2005). Expanding functional diversity of the coactivators. *Trends Biochem Sci*, *30*(3), 126-32.

Lopez-Maury, L., Marguerat, S., & Bahler, J. (2008). Tuning gene expression to changing environments : from rapid responses to evolutionary adaptation. *Nat Rev Genet*, *9*(8), 583-93.

Lopez Sanchez, A., Stassen, J. H., Furci, L., Smith, L. M., & Ton, J. (2016). The role of dna (de)methylation in immune responsiveness of arabidopsis. *Plant J*, *88*(3), 361-374.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, *389*(6648), 251-60.

Mallo, M., & Alonso, C. R. (2013). The regulation of hox gene expression during animal development. *Development*, *140*(19), 3951-63.

Malone, C. D., & Hannon, G. J. (2009). Small rnas as guardians of the genome. *Cell*, *136*(4), 656-68.

Marinus, M. G., & Lobner-Olesen, A. (2014). Dna methylation. *EcoSal Plus*, *6*(1).

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet*, *7*, 29-59.

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., Haussler, D., Marra, M. A., Hirst, M., Wang, T., & Costello, J. F. (2010). Conserved role of intragenic dna methylation in regulating alternative promoters. *Nature*, *466*(7303), 253-7.

Mayr, E. (1982). *The growth of biological thought diversity, evolution, and inheritance.* Cambridge, Massachusetts : Belknap Press, 1982.

Mazzio, E. A., & Soliman, K. F. (2012). Basic concepts of epigenetics : impact of environmental signals on gene expression. *Epigenetics*, *7*(2), 119-30.

McBride, C. S., Baier, F., Omondi, A. B., Spitzer, S. A., Lutomiah, J., Sang, R., Ignell, R., & Vosshall, L. B. (2014). Evolution of mosquito preference for humans linked to an odorant receptor. *Nature*, *515*, 222.

McCarroll, S. A., Murphy, C. T., Zou, S., Pletcher, S. D., Chin, C. S., Jan, Y. N., Kenyon, C., Bargmann, C. I., & Li, H. (2004). Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet*, *36*(2), 197-204.

McCulloch, S. D., & Kunkel, T. A. (2008). The fidelity of dna synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res*, *18*(1), 148-61.

McGhee, J. D., & Ginder, G. D. (1979). Specific dna methylation sites in the vicinity of the chicken beta-globin genes. *Nature*, *280*(5721), 419-20.

McRae, A. F., Powell, J. E., Henders, A. K., Bowdler, L., Hemani, G., Shah, S., Painter, J. N., Martin, N. G., Visscher, P. M., & Montgomery, G. W. (2014). Contribution of genetic variation to transgenerational inheritance of dna methylation. *Genome Biology*, *15*(5), R73.

Mele, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., Young, T. R., Goldmann, J. M., Pervouchine, D. D., Sullivan, T. J., Johnson, R., Segre, A. V., Djebali, S., Niarchou, A., Consortium, G. T., Wright, F. A., Lappalainen, T., Calvo, M., Getz, G., Dermitzakis, E. T., Ardlie, K. G., & Guigo, R. (2015). Human genomics. the human transcriptome across tissues and individuals. *Science*,

348(6235), 660-5.

Mendel, G. (1866). *Versuche über pflanzen-hybriden.* Brünn : : Im Verlage des Vereines.

Messerschmidt, D. M., Knowles, B. B., & Solter, D. (2014). Dna methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. *Genes Dev*, *28*(8), 812-28.

Michels, K. B., Binder, A. M., Dedeurwaerder, S., Epstein, C. B., Greally, J. M., Gut, I., Houseman, E. A., Izzi, B., Kelsey, K. T., Meissner, A., Milosavljevic, A., Siegmund, K. D., Bock, C., & Irizarry, R. A. (2013). Recommendations for the design and analysis of epigenome-wide association studies. *Nature Methods*, *10*, 949.

Moen, E. L., Zhang, X., Mu, W., Delaney, S. M., Wing, C., McQuade, J., Myers, J., Godley, L. A., Dolan, M. E., & Zhang, W. (2013). Genome-wide variation of cytosine modifications between european and african populations and the implications for complex traits. *Genetics*, *194*(4), 987-96.

Monks, S. A., Leonardson, A., Zhu, H., Cundiff, P., Pietrusiak, P., Edwards, S., Phillips, J. W., Sachs, A., & Schadt, E. E. (2004). Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet*, *75*(6), 1094-105.

Moore, L. D., Le, T., & Fan, G. (2013). Dna methylation and its basic function. *Neuropsychopharmacology*, *38*(1), 23-38.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., & Worm, B. (2011). How many species are there on earth and in the ocean ? *PLoS Biol*, *9*(8), e1001127.

Moran, S., Arribas, C., & Esteller, M. (2016). Validation of a dna methylation microarray for 850,000 cpg sites of the human genome enriched in enhancer sequences. *Epigenomics*, *8*(3), 389-99.

Nedelec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J. C., Freiman, A., Sams, A. J., Hebert, S., Page Sabourin, A., Luca, F., Blekhman, R., Hernandez, R. D., Pique-Regi, R., Tung, J., Yotova, V., & Barreiro, L. B. (2016). Genetic ancestry and natural selection drive population differences in immune responses to pathogens. *Cell*, *167*(3), 657-669 e21.

Netea, M. G., Joosten, L. A., Latz, E., Mills, K. H., Natoli, G., Stunnenberg, H. G., O'Neill, L. A., & Xavier, R. J. (2016). Trained immunity : A program of innate immune memory in health and disease. *Science*, *352*(6284), aaf1098.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated snps are more likely to be eqtls : annotation to enhance discovery from gwas. *PLoS Genet*, *6*(4), e1000888.

Nowick, K., Gernat, T., Almaas, E., & Stubbs, L. (2009). Differences in human and chimpanzee gene expression patterns define an evolving network of transcription factors in brain. *Proc Natl Acad Sci U S A*, *106*(52), 22358-63.

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microrna biogenesis, mechanisms of actions, and circulation. *Front Endocrinol (Lausanne)*, *9*, 402.

Okano, M., Bell, D. W., Haber, D. A., & Li, E. (1999). Dna methyltransferases dnmt3a and dnmt3b are essential for de novo methylation and mammalian development. *Cell*, *99*(3), 247-57.

Ollikainen, M., Smith, K. R., Joo, E. J., Ng, H. K., Andronikos, R., Novakovic, B., Abdul Aziz, N. K., Carlin, J. B., Morley, R., Saffery, R., & Craig, J. M. (2010). Dna methylation analysis of multiple tissues from newborn twins reveals both genetic and intrauterine components to variation in the human neonatal epigenome. *Hum Mol Genet*, *19*(21), 4176-88.

Ooi, S. K., Qiu, C., Bernstein, E., Li, K., Jia, D., Yang, Z., Erdjument-Bromage, H.,

Tempst, P., Lin, S. P., Allis, C. D., Cheng, X., & Bestor, T. H. (2007). Dnmt3l connects unmethylated lysine 4 of histone h3 to de novo methylation of dna. *Nature*, *448*(7154), 714-7.

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of rna polymerase ii. *Genes Dev*, *10*(21), 2657-83.

Ozata, D. M., Gainetdinov, I., Zoch, A., O'Carroll, D., & Zamore, P. D. (2019). Piwi-interacting rnas : small rnas with big functions. *Nat Rev Genet*, *20*(2), 89-108.

Pacis, A., Mailhot-Leonard, F., Tailleux, L., Randolph, H. E., Yotova, V., Dumaine, A., Grenier, J. C., & Barreiro, L. B. (2019). Gene activation precedes dna demethylation in response to infection in human dendritic cells. *Proc Natl Acad Sci U S A*, *116*(14), 6938-6943.

Pacis, A., Tailleux, L., Morin, A. M., Lambourne, J., MacIsaac, J. L., Yotova, V., Dumaine, A., Danckaert, A., Luca, F., Grenier, J. C., Hansen, K. D., Gicquel, B., Yu, M., Pai, A., He, C., Tung, J., Pastinen, T., Kobor, M. S., Pique-Regi, R., Gilad, Y., & Barreiro, L. B. (2015). Bacterial infection remodels the dna methylation landscape of human dendritic cells. *Genome Res*, *25*(12), 1801-11.

Pai, A. A., Bell, J. T., Marioni, J. C., Pritchard, J. K., & Gilad, Y. (2011). A genome-wide study of dna methylation patterns and gene expression levels in multiple human and chimpanzee tissues. *PLoS Genet*, *7*(2), e1001316.

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet*, *11*(1), e1004857.

Pardo-Diaz, C., Salazar, C., & Jiggins, C. D. (2015). Towards the identification of the loci of adaptive evolution. *Methods Ecol Evol*, *6*(4), 445-464.

Patin, E., Bergstedt, J., Ait Kaci Azzou, S., Urrutia, A., Quach, H., Tsuo, K., Husquin, L. T., Rotival, M., Kobor, M. S., Albert, M. L., Duffy, D., Quintana-Murci, L., & Consortium., f. t. M. I. (2019). Factors driving dna methylation variation in human blood. *(in preparation)*.

Patterson, K., Molloy, L., Qu, W., & Clark, S. (2011). Dna methylation : bisulphite modification and analysis. *J Vis Exp*(56).

Paul, S., & Amundson, S. A. (2014). Differential effect of active smoking on gene expression in male and female smokers. *J Carcinog Mutagen*, *5*.

Piasecka, B., Duffy, D., Urrutia, A., Quach, H., Patin, E., Posseme, C., Bergstedt, J., Charbit, B., Rouilly, V., MacPherson, C. R., Hasan, M., Albaud, B., Gentien, D., Fellay, J., Albert, M. L., Quintana-Murci, L., & Milieu Interieur, C. (2018). Distinctive roles of age, sex, and genetics in shaping transcriptional variation of human immune responses to microbial challenges. *Proc Natl Acad Sci U S A*, *115*(3), E488-E497.

Picascia, A., Grimaldi, V., Pignalosa, O., De Pascale, M. R., Schiano, C., & Napoli, C. (2015). Epigenetic control of autoimmune diseases : from bench to bedside. *Clin Immunol*, *157*(1), 1-15.

Plath, K., Mlynarczyk-Evans, S., Nusinow, D. A., & Panning, B. (2002). Xist rna and the mechanism of x chromosome inactivation. *Annu Rev Genet*, *36*, 233-78.

Pollack, Y., Stein, R., Razin, A., & Cedar, H. (1980). Methylation of foreign dna sequences in eukaryotic cells. *Proc Natl Acad Sci U S A*, *77*(11), 6463-7.

Popodi, E., Kissinger, J. C., Andrews, M. E., & Raff, R. A. (1996). Sea urchin hox genes : insights into the ancestral hox cluster. *Mol Biol Evol*, *13*(8), 1078-86.

Probst, A. V., Dunleavy, E., & Almouzni, G. (2009). Epigenetic inheritance during the cell cycle. *Nat Rev Mol Cell Biol*, *10*(3), 192-206.

Prud'homme, B., & Gompel, N. (2010). Genomic hourglass. *Nature*, *468*, 768.

Quach, H., Rotival, M., Pothlichet, J., Loh, Y. E., Dannemann, M., Zidane, N., Laval, G., Patin, E., Harmant, C., Lopez, M., Deschamps, M., Naffakh, N., Duffy, D., Coen, A., Leroux-Roels, G., Clement, F., Boland, A., Deleuze, J. F., Kelso, J., Albert, M. L., & Quintana-Murci, L. (2016). Genetic adaptation and neandertal admixture shaped the immune system of human populations. *Cell*, *167*(3), 643-656 e17.

Quinonez, S. C., & Innis, J. W. (2014). Human hox gene disorders. *Mol Genet Metab*, *111*(1), 4-15.

Qureshi, S. A., Bashir, M. U., & Yaqinuddin, A. (2010). Utility of dna methylation markers for diagnosing cancer. *International Journal of Surgery*, *8*(3), 194-198.

Raj, A., & van Oudenaarden, A. (2008). Nature, nurture, or chance : stochastic gene expression and its consequences. *Cell*, *135*(2), 216-26.

Rakyan, V. K., Down, T. A., Balding, D. J., & Beck, S. (2011). Epigenome-wide association studies for common human diseases. *Nature reviews. Genetics*, *12*(8), 529-541.

Ramirez-Carrozzi, V. R., Braas, D., Bhatt, D. M., Cheng, C. S., Hong, C., Doty, K. R., Black, J. C., Hoffmann, A., Carey, M., & Smale, S. T. (2009). A unifying model for the selective regulation of inducible transcription by cpg islands and nucleosome remodeling. *Cell*, *138*(1), 114-28.

Razin, A., & Cedar, H. (1991). Dna methylation and gene expression. *Microbiol Rev*, *55*(3), 451-8.

Reveron-Gomez, N., Gonzalez-Aguilera, C., Stewart-Morgan, K. R., Petryk, N., Flury, V., Graziano, S., Johansen, J. V., Jakobsen, J. S., Alabert, C., & Groth, A. (2018). Accurate recycling of parental histones reproduces the histone modification landscape during dna replication. *Mol Cell*, *72*(2), 239-249 e5.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y. C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K. H., Feizi, S., Karlic, R., Kim, A. R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L. H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T., & Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317-30.

Robinson, P. J., Fairall, L., Huynh, V. A., & Rhodes, D. (2006). Em measurements define the dimensions of the "30-nm" chromatin fiber : evidence for a compact, interdigitated structure. *Proc Natl Acad Sci U S A*, *103*(17), 6506-11.

Romero, I. G., Ruvinsky, I., & Gilad, Y. (2012). Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*, *13*(7), 505-16.

Rosier, F. (2012). *L'épigénétique, l'hérédité au-delà de l'adn.* Le Monde, Sciences et Techno.

Rothbart, S. B., Krajewski, K., Nady, N., Tempel, W., Xue, S., Badeaux, A. I., Barsyte-Lovejoy, D., Martinez, J. Y., Bedford, M. T., Fuchs, S. M., Arrowsmith, C. H., & Strahl, B. D. (2012). Association of uhrf1 with methylated h3k9 directs the maintenance of dna methylation. *Nat Struct Mol Biol*, *19*(11), 1155-60.

Russo, V. E. A., Martienssen, R. A., & Riggs, A. D. (1996). *Epigenetic mechanisms of gene regulation.* Plainview, N.Y. : Cold Spring Harbor Laboratory Press.

Saeed, S., Quintin, J., Kerstens, H. H., Rao, N. A., Aghajanirefah, A., Matarese, F., Cheng, S. C., Ratter, J., Berentsen, K., van der Ent, M. A., Sharifi, N., Janssen-Megens, E. M., Ter Huurne, M., Mandoli, A., van Schaik, T., Ng, A., Burden, F., Downes, K., Frontini, M., Kumar, V., Giamarellos-Bourboulis, E. J., Ouwehand, W. H., van der Meer, J. W., Joosten, L. A., Wijmenga, C., Martens, J. H., Xavier, R. J., Logie, C., Netea, M. G., & Stunnenberg, H. G. (2014). Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity. *Science*, *345*(6204), 1251086.

Saldanha, S. N., & Tollefsbol, T. O. (2018). Chapter 7 - epigenetic approaches to cancer therapy. In T. O. Tollefsbol (Ed.), *Epigenetics in human disease (second edition)* (Vol. 6, p. 219-247). Academic Press.

Salta, S., S, P. N., Fontes-Sousa, M., Lopes, P., Freitas, M., Caldas, M., Antunes, L., Castro, F., Antunes, P., Palma de Sousa, S., Henrique, R., & Jeronimo, C. (2018). A dna methylation-based test for breast cancer detection in circulating cell-free dna. *J Clin Med*, *7*(11).

Sato, F., Tsuchiya, S., Meltzer, S. J., & Shimizu, K. (2011). Micrornas and epigenetics. *FEBS J*, *278*(10), 1598-609.

Schaaper, R. M. (1993). Base selection, proofreading, and mismatch repair during dna replication in escherichia coli. *J Biol Chem*, *268*(32), 23762-5.

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. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, *6*(5), e107.

Scharer, C. D., Barwick, B. G., Youngblood, B. A., Ahmed, R., & Boss, J. M. (2013). Global dna methylation remodeling accompanies cd8 t cell effector function. *J Immunol*, *191*(6), 3419-29.

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res*, *22*(9), 1748-59.

Schneider, E., Pliushch, G., El Hajj, N., Galetzka, D., Puhl, A., Schorsch, M., Frauenknecht, K., Riepert, T., Tresch, A., Muller, A. M., Coerdt, W., Zechner, U., & Haaf, T. (2010). Spatial, temporal and interindividual epigenetic variation of functionally important dna methylation patterns. *Nucleic Acids Res*, *38*(12), 3880-90.

Schuettengruber, B., Bourbon, H. M., Di Croce, L., & Cavalli, G. (2017). Genome regulation by polycomb and trithorax : 70 years and counting. *Cell*, *171*(1), 34-57.

Scotland, R., Krell, F., & Valdecasas, A. (s. d.). *International institute for species exploration.*

Skinner, M. K., Manikkam, M., & Guerrero-Bosagna, C. (2010). Epigenetic transgenerational actions of environmental factors in disease etiology. *Trends Endocrinol Metab*, *21*(4), 214-22.

Smale, S. T. (2012). Transcriptional regulation in the innate immune system. *Curr Opin Immunol*, *24*(1), 51-7.

Smale, S. T., & Kadonaga, J. T. (2003). The rna polymerase ii core promoter. *Annu Rev Biochem*, *72*, 449-79.

Smallwood, S. A., & Kelsey, G. (2012). De novo dna methylation : a germ cell perspective. *Trends Genet*, *28*(1), 33-42.

Smith, Z. D., & Meissner, A. (2013). Dna methylation : roles in mammalian development. *Nat Rev Genet*, *14*(3), 204-20.

Spielman, R. S., Bastone, L. A., Burdick, J. T., Morley, M., Ewens, W. J., & Cheung, V. G. (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nat Genet*, *39*(2), 226-31.

Splinter, E., Heath, H., Kooren, J., Palstra, R. J., Klous, P., Grosveld, F., Galjart, N., & de Laat, W. (2006). Ctcf mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes Dev*, *20*(17), 2349-54.

Stein, R., Razin, A., & Cedar, H. (1982). In vitro methylation of the hamster adenine phosphoribosyltransferase gene inhibits its expression in mouse l cells. *Proc Natl Acad Sci U S A*, *79*(11), 3418-22.

Steinheimer, F. (2004). *Charles darwin's bird collection and ornithological knowledge during the voyage of h.m.s. "beagle", 1831–1836* (Vol. 145).

Storey, J. D., Madeoy, J., Strout, J. L., Wurfel, M., Ronald, J., & Akey, J. M. (2007). Gene-expression variation within and among human populations. *Am J Hum Genet*, *80*(3), 502-9.

Strahl, B. D., & Allis, C. D. (2000). The language of covalent histone modifications. *Nature*, *403*(6765), 41-5.

Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., Price, A., Raj, T., Nisbett, J., Nica, A. C., Beazley, C., Durbin, R., Deloukas, P., & Dermitzakis, E. T. (2012). Patterns of cis regulatory variation in diverse human populations. *PLoS Genet*, *8*(4), e1002639.

Suarez-Alvarez, B., Rodriguez, R. M., Fraga, M. F., & Lopez-Larrea, C. (2012). Dna methylation : a promising landscape for immune system-related diseases. *Trends Genet*, *28*(10), 506-14.

Sun, M., Kawamura, R., & Marko, J. F. (2011). Micromechanics of human mitotic chromosomes. *Phys Biol*, *8*(1), 015003.

Sun, Y. V. (2014). The influences of genetic and environmental factors on methylome-wide association studies for human diseases. *Curr Genet Med Rep*, *2*(4), 261-270.

Suzuki, M. M., & Bird, A. (2008). Dna methylation landscapes : provocative insights from epigenomics. *Nat Rev Genet*, *9*(6), 465-76.

Swalla, B. J. (2006). Building divergent body plans with similar genetic pathways. *Heredity (Edinb)*, *97*(3), 235-43.

Swanson-Wagner, R., Briskine, R., Schaefer, R., Hufford, M. B., Ross-Ibarra, J., Myers, C. L., Tiffin, P., & Springer, N. M. (2012). Reshaping of the maize transcriptome by domestication. *Proc Natl Acad Sci U S A*, *109*(29), 11878-83.

Szulwach, K. E., & Jin, P. (2014). Integrating dna methylation dynamics into a framework for understanding epigenetic codes. *Bioessays*, *36*(1), 107-17.

Tahiliani, M., Koh, K. P., Shen, Y., Pastor, W. A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L. M., Liu, D. R., Aravind, L., & Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian dna by mll partner tet1. *Science*, *324*(5929), 930-5.

Taylor, D. H., Chu, E. T., Spektor, R., & Soloway, P. D. (2015). Long non-coding rna regulation of reproduction and development. *Mol Reprod Dev*, *82*(12), 932-56.

Teschendorff, A. E., & Relton, C. L. (2018). Statistical and integrative system-level analysis of dna methylation data. *Nat Rev Genet*, *19*(3), 129-147.

Theissen, G. (2001). Development of floral organ identity : stories from the mads house. *Curr Opin Plant Biol*, *4*(1), 75-85.

Thomae, A. W., Schade, G. O., Padeken, J., Borath, M., Vetter, I., Kremmer, E., Heun, P., & Imhof, A. (2013). A pair of centromeric proteins mediates reproductive isolation in drosophila species. *Dev Cell*, *27*(4), 412-24.

Thu, K. L., Vucic, E. A., Kennett, J. Y., Heryet, C., Brown, C. J., Lam, W. L., & Wilson, I. M. (2009). Methylated dna immunoprecipitation. *J Vis Exp*(23).

Titus, A. J., Gallimore, R. M., Salas, L. A., & Christensen, B. C. (2017). Cell-type deconvolution from dna methylation : a review of recent applications. *Hum Mol Genet*, *26*(R2), R216-R224.

T Kalinka, A., M Varga, K., Gerrard, D., Preibisch, S., L Corcoran, D., Jarrells, J., Ohler, U., M Bergman, C., & Tomancak, P. (2010). *Gene expression divergence recapitulates the developmental hourglass model* (Vol. 468).

Toledo-Rodriguez, M., Lotfipour, S., Leonard, G., Perron, M., Richer, L., Veillette, S., Pausova, Z., & Paus, T. (2010). Maternal smoking during pregnancy is associated with epigenetic modifications of the brain-derived neurotrophic factor-6 exon in adolescent offspring. *Am J Med Genet B Neuropsychiatr Genet*, *153B*(7), 1350-4.

Tsaprouni, L. G., Yang, T. P., Bell, J., Dick, K. J., Kanoni, S., Nisbet, J., Vinuela, A., Grundberg, E., Nelson, C. P., Meduri, E., Buil, A., Cambien, F., Hengstenberg, C., Erdmann, J., Schunkert, H., Goodall, A. H., Ouwehand, W. H., Dermitzakis, E., Spector, T. D., Samani, N. J., & Deloukas, P. (2014). Cigarette smoking reduces dna methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. *Epigenetics*, *9*(10), 1382-96.

Tung, J., & Gilad, Y. (2013). Social environmental effects on gene regulation. *Cell Mol Life Sci*, *70*(22), 4323-39.

Unnikrishnan, A., Freeman, W. M., Jackson, J., Wren, J. D., Porter, H., & Richardson, A. (2019). The role of dna methylation in epigenetics of aging. *Pharmacol Ther*, *195*, 172-185.

Uszczynska-Ratajczak, B., Lagarde, J., Frankish, A., Guigo, R., & Johnson, R. (2018). Towards a complete map of the human long non-coding rna transcriptome. *Nat Rev Genet*, *19*(9), 535-548.

Valdelièvre, H., & Charraud, A. (1981). La taille et le poids des français. *Economie et Statistique*, 23-38.

Valinluck, V., Tsai, H. H., Rogstad, D. K., Burdzy, A., Bird, A., & Sowers, L. C. (2004). Oxidative damage to methyl-cpg sequences inhibits the binding of the methyl-cpg binding domain (mbd) of methyl-cpg binding protein 2 (mecp2). *Nucleic Acids Res*, *32*(14), 4100-8.

Vandiver, A. R., Irizarry, R. A., Hansen, K. D., Garza, L. A., Runarsson, A., Li, X., Chien, A. L., Wang, T. S., Leung, S. G., Kang, S., & Feinberg, A. P. (2015). Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant

skin. *Genome Biol*, *16*, 80.

van Eijk, K. R., de Jong, S., Boks, M. P., Langeveld, T., Colas, F., Veldink, J. H., de Kovel, C. G., Janson, E., Strengman, E., Langfelder, P., Kahn, R. S., van den Berg, L. H., Horvath, S., & Ophoff, R. A. (2012). Genetic analysis of dna methylation and gene expression levels in whole blood of healthy human subjects. *BMC Genomics*, *13*, 636.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304-51.

Vologodskii, A. V., & Cozzarelli, N. R. (1994). Conformational and thermodynamic properties of supercoiled dna. *Annu Rev Biophys Biomol Struct*, *23*, 609-43.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M. J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., Pervjakova, N., Alvaes, I., Fave, M.-J., Agbessi, M., Christiansen, M., Jansen, R., Seppälä, I., Tong, L., Teumer, A., Schramm, K., Hemani, G., Verlouw, J., Yaghootkar, H., Sönmez, R., Brown, A., Kukushkina, V., Kalnapenkis, A., Rüeger, S., Porcu, E., Kronberg-Guzman, J., Kettunen, J., Powell, J., Lee, B., Zhang, F., Arindrarto, W., Beutner, F., Brugge, H., Dmitreva, J., Elansary, M., Fairfax, B. P., Georges, M., Heijmans, B. T., Kähönen, M., Kim, Y., Knight, J. C., Kovacs, P., Krohn, K., Li, S., Loeffler, M., Marigorta, U. M., Mei, H., Momozawa, Y., Müller-Nurasyid, M., Nauck, M., Nivard, M., Penninx, B., Pritchard, J., Raitakari, O., Rotzchke, O., Slagboom, E. P., Stehouwer, C. D. A., Stumvoll, M., Sullivan, P., Hoen, P. A. C., Thiery, J., Tönjes, A., van Dongen, J., van Iterson, M., Veldink, J., Völker, U., Wijmenga, C., Swertz, M., Andiappan, A., Montgomery, G. W., Ripatti, S., Perola, M., Kutalik, Z., Dermitzakis, E., Bergmann, S., Frayling, T., van Meurs, J., Prokisch, H., Ahsan, H., Pierce, B., Lehtimäki, T., Boomsma, D., Psaty, B. M., Gharib, S. A., Awadalla, P., Milani, L., Ouwehand, W., Downes, K., Stegle, O., Battle, A., Yang, J., Visscher, P. M., Scholz, M., Gibson, G., Esko, T., & Franke, L. (2018). Unraveling the polygenic architecture of complex traits using blood eqtl metaanalysis. *bioRxiv*, 447367.

Waalwijk, C., & Flavell, R. A. (1978). Dna methylation at a ccgg sequence in the large intron of the rabbit beta-globin gene : tissue-specific variations. *Nucleic Acids Res*, *5*(12), 4631-4.

Waddington, C. H. (1940). *Organisers and genes*. Cambridge Biological Studies. University Press, Cambridge.

Waddington, C. H. (1957). *The strategy of the genes. a discussion of some aspects of theoretical biology. with an appendix by h. kacser.* London : George Allen & Unwin, Ltd.

Wagner, J. R., Busche, S., Ge, B., Kwan, T., Pastinen, T., & Blanchette, M. (2014). The relationship between dna methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biol*, *15*(2), R37.

Wahl, S., Drong, A., Lehne, B., Loh, M., Scott, W. R., Kunze, S., Tsai, P. C., Ried, J. S., Zhang, W., Yang, Y., Tan, S., Fiorito, G., Franke, L., Guarrera, S., Kasela, S., Kriebel, J., Richmond, R. C., Adamo, M., Afzal, U., Ala-Korpela, M., Albetti, B., Ammerpohl, O., Apperley, J. F., Beekman, M., Bertazzi, P. A., Black, S. L., Blancher, C., Bonder, M. J., Brosch, M., Carstensen-Kirberg, M., de Craen, A. J., de Lusignan, S., Dehghan, A., Elkalaawy, M., Fischer, K., Franco, O. H., Gaunt, T. R., Hampe, J., Hashemi, M., Isaacs, A., Jenkinson, A., Jha, S., Kato, N., Krogh, V., Laffan, M., Meisinger, C., Meitinger, T., Mok, Z. Y., Motta, V., Ng, H. K., Nikolakopoulou, Z., Nteliopoulos, G., Panico, S., Pervjakova, N., Prokisch, H., Rathmann, W., Roden, M., Rota, F., Rozario, M. A., Sandling, J. K., Schafmayer, C., Schramm, K., Siebert, R., Slagboom, P. E., Soininen, P., Stolk, L., Strauch, K., Tai, E. S., Tarantini, L., Thorand, B., Tigchelaar, E. F., Tumino, R., Uitterlinden, A. G., van Duijn, C., van Meurs, J. B., Vineis, P., Wickremasinghe, A. R., Wijmenga, C., Yang, T. P., Yuan, W., Zhernakova, A., Batterham, R. L., Smith, G. D., Deloukas, P., Heijmans, B. T., Herder, C., Hofman, A., Lindgren, C. M., Milani, L., van der Harst, P., Peters, A., Illig, T., Relton, C. L., Waldenberger, M., Jarvelin, M. R., Bollati, V., Soong, R., Spector, T. D., Scott, J., McCarthy, M. I., et al. (2017). Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, *541*(7635), 81-86.

Wang, J., Hevi, S., Kurash, J. K., Lei, H., Gay, F., Bajko, J., Su, H., Sun, W., Chang, H., Xu, G., Gaudet, F., Li, E., & Chen, T. (2009). The lysine demethylase lsd1 (kdm1) is required for maintenance of global dna methylation. *Nat Genet*, *41*(1), 125-9.

Wang, L., Pittman, K. J., Barker, J. R., Salinas, R. E., Stanaway, I. B., Williams, G. D., Carroll, R. J., Balmat, T., Ingham, A., Gopalakrishnan, A. M., Gibbs, K. D., Antonia, A. L., e, M. N., Heitman, J., Lee, S. C., Jarvik, G. P., Denny, J. C., Horner, S. M., DeLong, M. R., Valdivia, R. H., Crosslin, D. R., & Ko, D. C. (2018). An atlas of genetic variation linking pathogen-induced cellular traits to human disease. *Cell Host Microbe*, *24*(2), 308-323 e6.

Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nucleic acids : A structure for deoxyribose nucleic acid. *Nature*, *171*(4356), 737-738.

Wei, Y., Yang, C. R., Wei, Y. P., Zhao, Z. A., Hou, Y., Schatten, H., & Sun, Q. Y. (2014). Paternally induced transgenerational inheritance of susceptibility to diabetes in mammals. *Proc Natl Acad Sci U S A*, *111*(5), 1873-8.

Weidner, C. I., Lin, Q., Koch, C. M., Eisele, L., Beier, F., Ziegler, P., Bauerschlag, D. O., Jöckel, K.-H., Erbel, R., Mühleisen, T. W., Zenke, M., Brümmendorf, T. H., & Wagner, W. (2014). Aging of blood can be tracked by dna methylation changes at just three cpg sites. *Genome Biology*, *15*(2), R24.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., & Parkinson, H. (2014). The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Res*, *42*(Database issue), D1001-6.

Wigler, M., Levy, D., & Perucho, M. (1981). The somatic replication of dna methylation.

Cell, *24*(1), 33-40.

Wolffe, A. (1998). *Chromatin : structure and function* (3rd éd.). San Diego : Academic Press.

Woodcock, C. L. (2006). Chromatin architecture. *Curr Opin Struct Biol*, *16*(2), 213-20.

Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet*, *8*(3), 206-16.

Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V., & Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol*, *20*(9), 1377-419.

Wu, H., & Zhang, Y. (2014). Reversing dna methylation : mechanisms, genomics, and biological functions. *Cell*, *156*(1-2), 45-68.

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., Liu, Y., Byrum, S. D., Mackintosh, S. G., Zhong, M., Tackett, A., Wang, G., Hon, L. S., Fang, G., Swenberg, J. A., & Xiao, A. Z. (2016). Dna methylation on n(6)-adenine in mammalian embryonic stem cells. *Nature*, *532*(7599), 329-33.

Xu, Q., Xing, S., Zhu, C., Liu, W., Fan, Y., Wang, Q., Song, Z., Yang, W., Luo, F., Shang, F., Kang, L., Chen, W., Yan, J., Li, J., & Sang, T. (2015). Population transcriptomics reveals a potentially positive role of expression diversity in adaptation. *J Integr Plant Biol*, *57*(3), 284-99.

Yan, H., Yuan, W., Velculescu, V. E., Vogelstein, B., & Kinzler, K. W. (2002). Allelic variation in human gene expression. *Science*, *297*(5584), 1143.

Yang, C., Khanniche, A., DiSpirito, J. R., Ji, P., Wang, S., Wang, Y., & Shen, H. (2016). Transcriptome signatures reveal rapid induction of immune-responsive genes in human memory cd8(+) t cells. *Sci Rep*, *6*, 27005.

Yang, X., Han, H., De Carvalho, D. D., Lay, F. D., Jones, P. A., & Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, *26*(4), 577-90.

Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P. K., Kivioja, T., Dave, K., Zhong, F., Nitta, K. R., Taipale, M., Popov, A., Ginno, P. A., Domcke, S., Yan, J., Schubeler, D., Vinson, C., & Taipale, J. (2017). Impact of cytosine methylation on dna binding specificities of human transcription factors. *Science*, *356*(6337).

Young, T., Rowland, J. E., van de Ven, C., Bialecka, M., Novoa, A., Carapuco, M., van Nes, J., de Graaff, W., Duluc, I., Freund, J. N., Beck, F., Mallo, M., & Deschamps, J. (2009). Cdx and hox genes differentially regulate posterior axial growth in mammalian embryos. *Dev Cell*, *17*(4), 516-26.

Yuan, T., Jiao, Y., de Jong, S., Ophoff, R. A., Beck, S., & Teschendorff, A. E. (2015). An integrative multi-scale analysis of the dynamic dna methylation landscape in aging. *PLoS Genet*, *11*(2), e1004996.

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556-9.

Zaret, K. S., & Carroll, J. S. (2011). Pioneer transcription factors : establishing competence for gene expression. *Genes Dev*, *25*(21), 2227-41.

Zhang, W., Duan, S., Kistner, E. O., Bleibel, W. K., Huang, R. S., Clark, T. A., Chen, T. X., Schweitzer, A. C., Blume, J. E., Cox, N. J., & Dolan, M. E. (2008). Evaluation of genetic variation contributing to differences in gene expression between populations. *Am J Hum Genet*, *82*(3), 631-40.

Zhang, X., Hu, M., Lyu, X., Li, C., Thannickal, V. J., & Sanders, Y. Y. (2017). Dna methylation regulated gene expression in organ fibrosis. *Biochim Biophys Acta Mol Basis Dis*, *1863*(9), 2389-2397.

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., & Ecker, J. R. (2006). Genome-wide high-resolution mapping and functional analysis of dna methylation in arabidopsis. *Cell*, *126*(6), 1189-201.

Zhong, H., Beaulaurier, J., Lum, P. Y., Molony, C., Yang, X., Macneil, D. J., Weingarth, D. T., Zhang, B., Greenawalt, D., Dobrin, R., Hao, K., Woo, S., Fabre-Suver, C., Qian, S., Tota, M. R., Keller, M. P., Kendziorski, C. M., Yandell, B. S., Castro, V., Attie, A. D., Kaplan, L. M., & Schadt, E. E. (2010). Liver and adipose expression associated snps are enriched for association to type 2 diabetes. *PLoS Genet*, *6*(5), e1000932.

Zhu, H., Wang, G., & Qian, J. (2016). Transcription factors as readers and effectors of dna methylation. *Nat Rev Genet*, *17*(9), 551-65.

Zilberman, D., Coleman-Derr, D., Ballinger, T., & Henikoff, S. (2008). Histone h2a.z and dna methylation are mutually antagonistic chromatin marks. *Nature*, *456*(7218), 125-9.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., & Henikoff, S. (2007). Genome-wide analysis of arabidopsis thaliana dna methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, *39*(1), 61-9.

Ziller, M. J., Gu, H., Muller, F., Donaghey, J., Tsai, L. T., Kohlbacher, O., De Jager, P. L., Rosen, E. D., Bennett, D. A., Bernstein, B. E., Gnirke, A., & Meissner, A. (2013). Charting a dynamic dna methylation landscape of the human genome. *Nature*, *500*(7463), 477-81.

Annexe 1 : Évaluation des méthodes d'estimation de l'âge par la méthylation de l'ADN après différentes méthodes de normalisation sur la puce Infinium MethylationEPIC BeadChip.

McEwen et al. Clinical Epigenetics (2018) 10:123
https://doi.org/10.1186/s13148-018-0556-2

Clinical Epigenetics

SHORT REPORT Open Access

Systematic evaluation of DNA methylation age estimation with common preprocessing methods and the Infinium MethylationEPIC BeadChip array

Lisa M McEwen[1*], Meaghan J Jones[1], David Tse Shen Lin[1], Rachel D Edgar[1], Lucas T Husquin[3,4,5], Julia L MacIsaac[1], Katia E Ramadori[1], Alexander M Morin[1], Christopher F Rider[2], Chris Carlsten[2], Lluís Quintana-Murci[3,4,5], Steve Horvath[6] and Michael S Kobor[1]

Abstract

Background: The capacity of technologies measuring DNA methylation (DNAm) is rapidly evolving, as are the options for applicable bioinformatics methods. The most commonly used DNAm microarray, the Illumina Infinium HumanMethylation450 (450K array), has recently been replaced by the Illumina Infinium HumanMethylationEPIC (EPIC array), nearly doubling the number of targeted CpG sites. Given that a subset of 450K CpG sites is absent on the EPIC array and that several tools for both data normalization and analyses were developed on the 450K array, it is important to assess their utility when applied to EPIC array data. One of the most commonly used 450K tools is the pan-tissue epigenetic clock, a multivariate predictor of biological age based on DNAm at 353 CpG sites. Of these CpGs, 19 are missing from the EPIC array, thus raising the question of whether EPIC data can be used to accurately estimate DNAm age. We also investigated a 71-CpG epigenetic age predictor, referred to as the Hannum method, which lacks 6 probes on the EPIC array. To evaluate these epigenetic clocks in EPIC data properly, a prior assessment of the effects of data preprocessing methods on DNAm age is also required.

Methods: DNAm was quantified, on both the 450K and EPIC platforms, from human primary monocytes derived from 172 individuals. We calculated DNAm age from raw, and three different preprocessed data forms to assess the effects of different processing methods on the DNAm age estimate. Using an additional cohort, we also investigated DNAm age of peripheral blood mononuclear cells, bronchoalveolar lavage, and bronchial brushing samples using the EPIC array.

Results: Using monocyte-derived data from subjects on both the 450K and EPIC, we found that DNAm age was highly correlated across both raw and preprocessing methods ($r > 0.91$). Thus, the correlation between chronological age and the DNAm age estimate is largely unaffected by platform differences and normalization methods. However, we found that the choice of normalization method and measurement platform can lead to a systematic offset in the age estimate which in turn leads to an increase in the median error. Comparing the 450K and EPIC DNAm age estimates, we observed that the median absolute difference was 1.44–3.10 years across preprocessing methods.

(Continued on next page)

* Correspondence: lmcewen@bcchr.ca
[1]BC Children's Hospital Research Institute, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, TRB A5-151, Vancouver, BC V5Z 4H4, Canada
Full list of author information is available at the end of the article

(Continued from previous page)
Conclusions: Here, we have provided evidence that the epigenetic clock is resistant to the lack of 19 CpG sites missing from the EPIC array as well as highlighted the importance of considering the technical variance of the epigenetic when interpreting group differences below the reported error. Furthermore, our study highlights the utility of epigenetic age acceleration measure, the residuals from a linear regression of DNAm age on chronological age, as the resulting values are robust with respect to normalization methods and measurement platforms.

Keywords: Epigenetic age, DNA methylation age, Epigenetic clock, EPIC, DNA methylation, 450K, Human, Microarray, Preprocessing

Background

Epigenetics is a rapidly evolving field in the contexts of new biological discoveries as well as the available technologies used to drive such findings. The most commonly studied epigenetic mark in humans is DNA methylation (DNAm), defined as the covalent addition of a methyl group to DNA, most frequently occurring at cytosine-guanine dinucleotides (CpGs) [20]. DNAm profiles change naturally during the development of an organism, resulting in tissue identity being the strongest predictor of DNAm variation. As such, DNAm variability between tissues, for example, the blood and brain, within an individual can be larger than the variability observed across individuals from the same tissue [7, 27]. Inter-individual variability in DNAm has been linked to a number of different sources, including but not limited to the underlying DNA sequence, environmental exposures, and health outcomes. One of the most active areas of research related to DNAm inter-individual variability in human cohorts focuses on the relationship between DNAm and aging, as there has been substantial evidence that DNAm changes with age, both linearly and non-linearly, across the entire life course [12].

Rapidly evolving new technologies and resources have fueled exponential growth in human DNAm research over the past decade, enhancing our ability to address questions such as the effects of aging. Although many methodologies can be used to measure DNAm, Illumina microarrays are the most common method for population-based epigenetic studies, as they provide an economical and accessible high-throughput platform. Over a little more than a decade, the capacity of Illumina DNAm microarray platform has increased from 1506 CpGs to more than 860,000 CpGs. The increased numbers of CpGs reflect both better coverage across genes and expanded interrogation of genomic regions. For example, the Illumina 27 K (27 K) array targeted > 27,000 CpG sites and interrogated at least one CpG per gene, but was biased towards CpG islands [2]. Its successor, the Illumina Infinium Methylation450 (450K) array assessed > 485,000 CpGs and covered 99% of RefSeq genes. The Illumina Infinium MethylationEPIC (EPIC) array is the newest tool and allows the quantification of over 860,000 CpG sites, with the additional content providing higher coverage of specific genomic regions, such as enhancers and non-coding regions. The EPIC array generally uses the same DNAm measurement protocol as the 450K array and includes over 94% of the 450K content [24]. However, the increased genomic resolution and complexity of the EPIC array in conjunction with missing 6% of the 450K CpGs necessitates an evaluation of the applicability of established bioinformatic tools established for the 27 K or 450K arrays.

To accommodate advancements in DNAm array technology and the increasing volume of data, many pipelines for data preprocessing, normalization, and analyses have been developed to streamline data handling [1, 5, 25, 28, 30]. Here, we refer to "preprocessing methods" as algorithms commonly performed on DNAm data prior to probe-type normalization, including methods to reduce background fluorescence or adjust for dye bias, which if unaddressed can reduce the dynamic range of beta values [31]. Probe-type normalization is a necessary adjustment for Illumina microarray DNAm data, as there are two different probe designs that possess differential beta distributions [2]. Tools such as the R function 'preprocessNoob' in the minfi package subtract background based on the out-of-band intensities (for example, Infinium I probes fluorescing in the color channel opposite their designed base extension). Color or dye bias adjustment is applied to account for the two color channels that type II probes employ, one for methylated and one for unmethylated CpGs, since residual dye can introduce unwanted variation. Tools to account for the color bias include in the Bioconductor package 'methylumi,' which is based on smooth quantile normalization, or the Illumina GenomeStudio software which implements a shift-and-scaling normalization [6]. Although these methods have been reviewed in comparison to one another [33, 35, 36], a mixed variety of pipelines are used across the literature and the influence of method selection on detecting true positives or generating accurate predictions should be investigated both within and across array technologies.

One tool that could be compromised by different preprocessing methods or the lack of certain 450K CpG sites on the EPIC array is the pan-tissue epigenetic clock, a

popular predictive model that estimates an individual's biological age, irrespective of tissue type, using DNAm at 353 CpGs [12]. Established on DNAm profiles (obtained from 27 K and 450K data) from 51 different tissues from over 8000 individuals, the epigenetic clock calculates DNAm age, which has been shown to correlate well with chronological age ($r > 0.80$) across the life course [14]. This epigenetic clock is hypothesized to be an accurate molecular biomarker of biological aging and deviations between chronological and DNAm age, commonly referred to as epigenetic age acceleration (which can be positive or negative), have been correlated with a host of age-related conditions, such as Parkinson's disease, time until death, frailty, and cognitive and physical decline [4, 15, 22, 23].

There are several other DNAm-based age predictors that have been reported [3, 9, 34], but another commonly used age predictor, specific to blood samples and referred to as the Hannum clock, is based on methylation at 71 CpG sites has also been observed to predict age with impressive accuracy. However, both the Horvath and Hannum models are lacking CpG sites on the on the EPIC array (19 of the 353 pan-tissue clock-CpGs and 6/71 of the Hannum clock CpGs are missing), and since the 450K platform is no longer available, it is crucial to assess the performance of these tools despite the missing probes, if use is to be continued.

Here, we investigated (1) the consistency between DNAm age measured from 450K and EPIC array data from the same individuals to evaluate the utility of EPIC array data given that it is missing clock CpGs used in the Horvath and Hannum age predictors, and (2) whether DNAm age estimates differ with different preprocessing methods. We found that EPIC data can be used to predict DNAm age accurately using both assessed epigenetic clocks. Additionally, we observed differences in DNAm age across preprocessing methods, although the differences across the values were below the reported median absolute error of the epigenetic clock. Lastly, we have replicated accurate measurement of DNAm age, using the pan-tissue predictor, across tissues using an EPIC dataset with three different tissues from 13 individuals. Our findings support the epigenetic clock as a robust tool that may be applied with EPIC array data in the future.

Methods

Cohort characteristics

We used two different cohorts in order to assess the pan-tissue epigenetic clock on the EPIC array. The first consisted of primary monocytes collected from 172 healthy males, aged 19–50 years old, of self-reported African- and European-descent from the EVOIMMUNOPOP project [19]. Genomic DNA was isolated from the monocyte fraction using a phenol/chloroform protocol followed by ethanol precipitation, and then subjected to bisulfite conversion with the EZ DNA Methylation Kit (ZymoResearch, Irvine, CA, USA). We quantified DNAm on all samples using two separate Illumina microarray platforms: 450K and EPIC arrays (Illumina, San Diego, CA, USA), following the manufacturer's instructions. To ensure sample labeling across technologies, we assessed the correlation between the overlapping quality control single-nucleotide polymorphic (SNP) probes present on both microarrays (59 SNPs); observing all sample pairs correlated with a Pearson's coefficient of $r \geq 0.99$ (Additional file 1: Figure S1). Four technical replicates were included during the 450K processing and 12 technical replicates were included during the EPIC sample processing, with two common technical replicates across technologies.

A secondary cohort was used to investigate EPIC DNA methylation data derived from tissues other than the blood. This cohort consisted of 13 individuals aged 23–46 years old from the control subset of Diesel Exhaust Study III (DE3) and was comprised of DNAm from peripheral blood mononuclear cells (PBMCs), bronchoalveolar lavage (BAL), and bronchial brushings (brush). All samples were collected from individuals after control (filtered air/saline) exposures. Primary cohort characteristics are provided in Additional file 1: Table S1. Note, approximately half of the individuals had prior physician-diagnosed asthma. Genomic DNA was isolated using the DNeasy Blood and Tissue Kit (Qiagen, Hilden Germany) and subsequently bisulfite converted using the EZ DNA Methylation Kit (ZymoResearch, Irvine, CA, USA). Bisulfite-treated samples were processed using the EPIC array as above (Illumina, San Diego, CA, USA).

DNA methylation quantification

All microarrays were scanned with an Illumina HiScan system. For the EPIC array data, we used the most current manifest file, "Infinium MethylationEPIC v1.0 B4 Manifest File," released by Illumina on May 26, 2017 and consisting of 865,918 probes, whereas for the 450K we used the "HumanMethylation450 v1.2 Manifest File" with 485,577 probes. Both manifest files are available at https://support.illumina.com/downloads.html. In addition to unprocessed (raw) data, we used data preprocessed in three different ways (1) color corrected/background subtracted in Genome Studio (GS), (2) quantile-normalized using "preprocessQuantile" [29], (3) normal-expontential out-of-band (noob)-normalized with "preprocessNoob" [30]. Raw data and data that were to be quantile or noob normalized were uploaded directly into R from IDAT files using the 'minfi' package function "read.metharray" [8]. For the color correction/background subtracted preprocessing, data were background subtracted/color corrected with GenomeStudio, and then uploaded into R with the package 'methylumi,' function 'lumiMethyR' [5].

DNA methylation age

We calculated DNAm age for each sample by using a modified version of the publicly available R code at https://dnamage.genetics.ucla.edu, with the normalization feature set to "TRUE" [12]. We focused our inquiry on data preprocessing only, and not probe-type normalization methods, as the epigenetic clock code applies an imputation of missing values and performs a calibrated version of a beta-mixture quantile normalization [12, 28]. The 71-CpG Hannum method age estimates were generated using methods described previously [10].

Results

The epigenetic clock accurately predicted DNA methylation age from EPIC methylation data

From the 450K array, 33,059 of 485,557 (6.8%) of probes are not represented on the EPIC array, including 19/353 epigenetic clock-CpGs (5.4%). The lack of 19 epigenetic clock CpGs on the EPIC array could reduce the accuracy of the epigenetic clock when using EPIC array data. Therefore, we investigated the consistency between DNAm age as calculated from the 450K (original 353-CpG model) and the EPIC (reduced 334-CpG model) arrays.

Focusing on a recently published data set of DNAm in purified monocytes from the EVOIMMUNOPOP project [19], we applied the epigenetic clock to data from samples run on both platforms and found a high correlation between the 450K and EPIC array DNAm age values regardless of preprocessing method ($r = 0.91–0.96$, error = 1.44–3.10 years, $R^2 = 0.83–0.91$, Fig. 1), observing consistent patterns between chronological age and DNAm age as measured from both the EPIC ($r = 0.84–0.86$) and 450K ($r = 0.86–0.87$) arrays (Additional file 1: Figure S2). Additionally, we performed probe-wise correlations of log transformed beta values at the 334 common clock CpG sites across the two platforms and found a range of Pearson's correlation coefficients $r = -0.18–0.98$ (Additional file 1: Figure S3A); specifically, out of the 334 probes an average (across preprocessing data sets) of 146 (44%) had ≤ 0.20, 118 (35%) had $r > 0.50$, and 44 (13%) had $r > 0.80$. Previous reports showed low correlation associated with low variation across the EPIC and 450K arrays, and so we tested to determine whether the clock probes with low correlation were also invariable. We

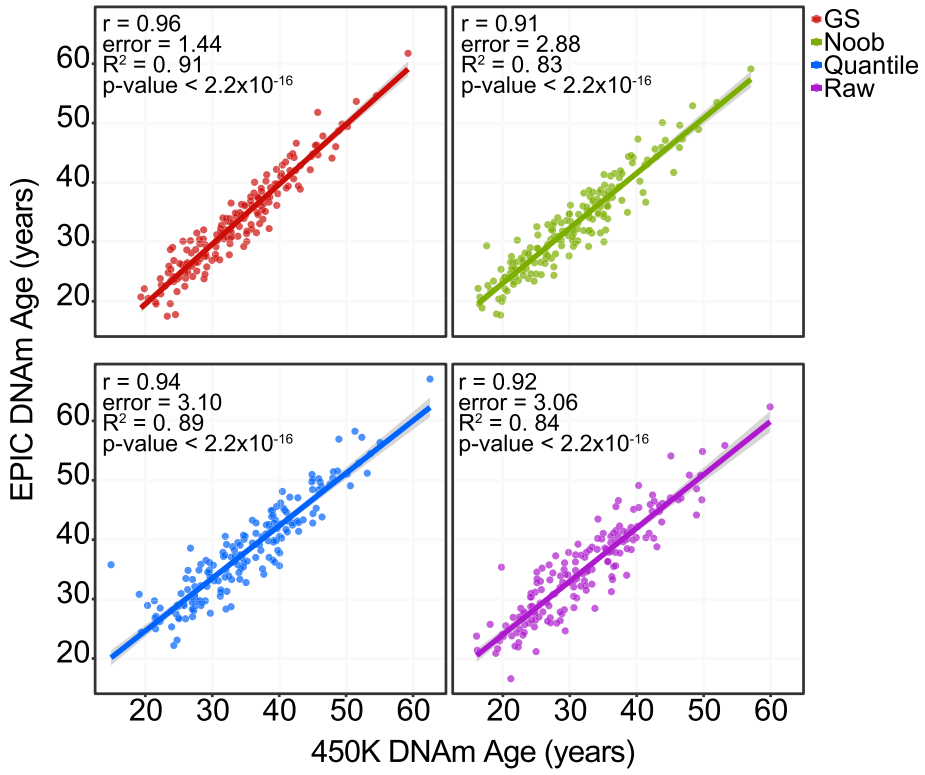

Fig. 1 DNA methylation age comparison between 450K or EPIC Monocyte data across preprocessing methods. Identical samples were assayed on both the 450K and EPIC arrays, and then each preprocessed in one of four ways prior to calculating DNA methylation (DNAm) age: raw unprocessed, GenomeStudio color correction/background subtraction (GS), normal exponential out-of-band (noob) normalization, or quantile normalization. Solid colored line represents corresponding group regression line. For each regression, the Pearson's correlation coefficient, error (median absolute error between EPIC DNAm age and 450K DNAm age), R^2 value, and p value corresponding to the correlation coefficient are shown

observed that lower beta value ranges were strongly associated with lower correlation values between the EPIC and 450K ($r = 0.75–78$, depending on preprocessing method, Additional file 1: Figure S3B). As a second approach to assess the direct consequence of the absent clock sites, we removed the 19 missing EPIC clock-CpGs in the 450K data to simulate the 334-CpG model. We then calculated DNAm age using both the 353-CpG model and the 334-CpG model from the 450k data, finding a strong correlation of $r = 0.998$, indicating that the missing 19 CpGs did not adversely affect DNAm age prediction in monocytes (Additional file 1: Figure S4). Furthermore, we calculated another DNAm age measure based on the Hannum method using 71 CpG sites of which only 6 (8.5%) are missing on the EPIC array [10], and again found strong correlations between Hannum DNAm age as calculated from EPIC and 450K array data ($r = 0.92–0.95$, Additional file 1: Figure S5).

Lastly, to confirm that the epigenetic clock could produce an accurate estimate of age from EPIC data across different tissues, we used an independent cohort of three tissues (PBMCs, BALs, brushes) collected from 13 healthy adults. Importantly, data from lung tissues has been reported previously to accurately estimate DNAm age in the context of DNAm age using the 450K array [12]. Calculating DNAm age with GS-preprocessed data, we observed a strong correlation with chronological age using EPIC data for the PBMC and BAL samples ($r = 0.88$, $r = 0.89$, respectively), but a lesser degree of correlation for the brush samples ($r = 0.59$, Fig. 2). We note that the brush beta-value distribution appeared to have higher inter-individual variability than the other tissues, which may explain the lower correlation in brush samples (Additional file 1: Figure S6).

Data preprocessing methods affected the calculated DNA methylation age, but within error margins of the epigenetic clock

Given that there is not an accepted standard practiced method of preprocessing data prior to calculating DNAm age, we assessed the potential effects of different commonly used data preprocessing methods on the DNAm age estimates. We compared DNAm age estimates calculated from raw data as well as after applying three separate standard data preprocessing methods: color-correction and background-normalization with GenomeStudio software (abbreviated GS), quantile-normalization, or noob-normalization. Imputation and a probe-type normalization were performed the same way across preprocessing methods using the R code supplied with the epigenetic clock method [16, 28, 32]. Using monocyte-derived data from 172 subjects on both the 450K and EPIC, we found that DNAm age was highly correlated across both raw data and data after three different preprocessing methods ($r > 0.91$) (Additional file 1: Figure S2). However, shifts in mean DNAm age were observed, indicating that although mean DNAm differences did exist, the trends with age were consistent across preprocessing methods (Fig. 3a). This was further supported by significant Kendall rank coefficients in DNAm age across each preprocessing method ($\tau = 0.86–0.94$, p value $< 2.2 \times 10^{-6}$, Additional file 1: Figure S7).

To further investigate the sample-to-sample trend in DNAm age across methods, we explored two common measures associated with the epigenetic clock, both considered measures of epigenetic age acceleration; the difference between DNAm age and chronological age (age acceleration difference) and the residuals from a linear model of DNAm age regressed onto chronological age

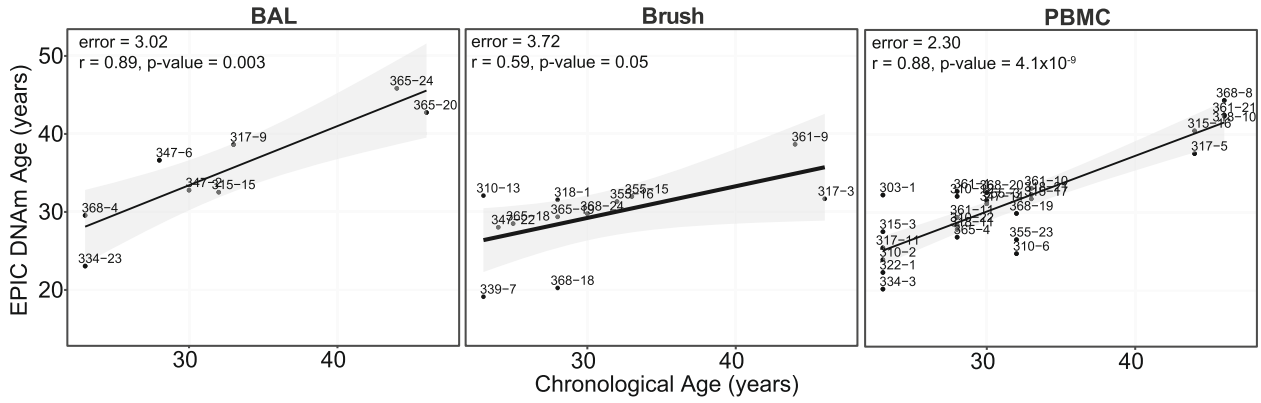

Fig. 2 EPIC DNA methylation age estimated in control samples from the Diesel Exhaust III Study across three tissues. DNA methylation (DNAm) age was estimated using the EPIC 334-CpG model from GenomeStudio background-subtracted and color-channel-adjusted EPIC data. Linear regression line shown with 95% confidence intervals is shown in gray. Error is the median absolute difference between EPIC DNAm age and chronological age. Pearson's correlation coefficients (r) and corresponding p value are shown for each tissue. BAL = bronchoalveolar lavage, PBMC = peripheral blood mononuclear cells, brush = bronchial brushing

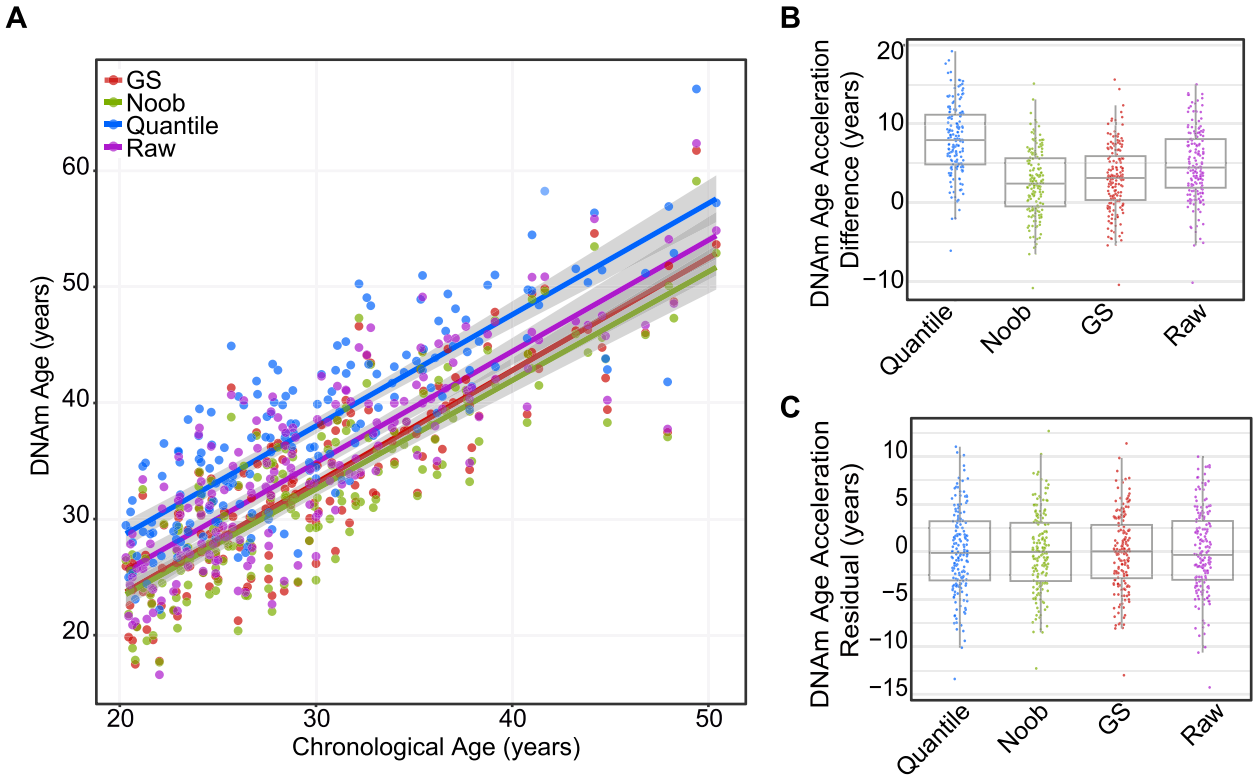

Fig. 3 DNA methylation age acceleration variation across preprocessing methods. **a** Scatter plot of EPIC DNA methylation (DNAm) age calculated from raw and data from three different preprocessing methods: quantile, GenomeStudio (GS), and normal exponential out-of-band (noob) normalization. Colored regression lines and surrounding shaded gray areas represent 95% confidence interval for each group. **b** Boxplot of estimated DNA methylation (DNAm) age—chronological age (acceleration difference) for each preprocessing method. **c** Boxplot of residuals from a linear regression (DNAm age~chronological age) across methods. The median is illustrated by horizontal line with upper and lower hinges representing the 25th and 75th percentiles, upper and lower whiskers extend no further than the inter-quartile range multiplied by 1.5. Colored data points represent individual samples for each group

(age acceleration residual). Since the observed mean DNAm age shifts when using different preprocessing methods (Fig. 3a), age acceleration difference is more likely to be affected by which preprocessing method was chosen. In contrast, age acceleration residual is less affected by mean differences as it is expressed relative to the measured population. As expected, we observed significant discrepancies in the mean age acceleration difference measure for nearly all comparisons (p value < 0.0002 for all age acceleration difference comparisons except for noob versus GS p value $= 0.23$, median absolute difference ranging from 0.68–5.55 years (Fig. 3b, Additional file 1: Table S2). Minimal variation was observed for the age acceleration residual mean across preprocessing methods (p value > 0.99, median absolute difference ranging from 0.52–1.23 years, Fig. 3c, Additional file 1: Table S2). This supports the previous suggestion of using age-acceleration residuals [12] to correct for processing specific shifts in DNAm estimates in order to accurately compare DNAm age between people.

We assessed how different preprocessing methods influenced the DNAm age estimate by examining the concordance of DNAm age measured from EPIC array technical replicates. A technical replicate pair represented an identical DNA sample quantified twice for quality control purposes; specifically, the sample was divided into two separate tubes after bisulfite conversion and DNAm was quantified separately. Technical replicate sample identity was confirmed by examining the 59 SNP probes present on the EPIC array (Additional file 1: Figure S8). We focused on the 24 technical replicates (12 pairs) from the EPIC array, calculating DNAm age for each technical replicate from data subjected to each separate preprocessing method: raw, GS color corrected and background subtracted, quantile normalized, or noob normalized. We calculated the median absolute difference between each technical replicate pair's DNAm age estimates in each dataset. We found that the GS color correction and background subtraction had the least deviation across replicates ($error_{GS} = 2.17$ years), followed by noob normalization ($error_{Noob} = 2.41$ years),

quantile normalization ($error_{Quantile}$ = 2.89 years), and then raw data having the largest deviation ($error_{Raw}$ = 3.14 years, Fig. 4). Notably, these values are all below the median absolute error of the epigenetic clock (3.6 years) [12].

Discussion

This study had two primary aims (1) to investigate whether using EPIC methylation data to calculate DNAm age is an appropriate approach, given the 19 and 6 missing probes used in the pan tissue 353 CpG model (Horvath) and 71 CpG model (Hannum), respectively, and (2) to evaluate the effect of various data preprocessing methods prior to calculating DNAm age, as a standard pipeline for processing data prior to calculating DNAm age does not exist. By analyzing monocyte DNAm from 172 individuals quantified on both the 450K and EPIC arrays, we demonstrated that the lack of the clock-CpGs on the EPIC array did not compromise the utility of the epigenetic age predictors. We also evaluated the performance of the EPIC pan-tissue epigenetic clock (334-CpG model) on another EPIC dataset, consisting of three tissues from 13 individuals, finding comparable correlations with age to those reported with the 450K array (353-CpG model). Furthermore, we found small differences in the DNAm age estimate between data preprocessing methods, implying that although the methods assessed here differed in mean values, the trends in respect to chronological age were consistent across methods.

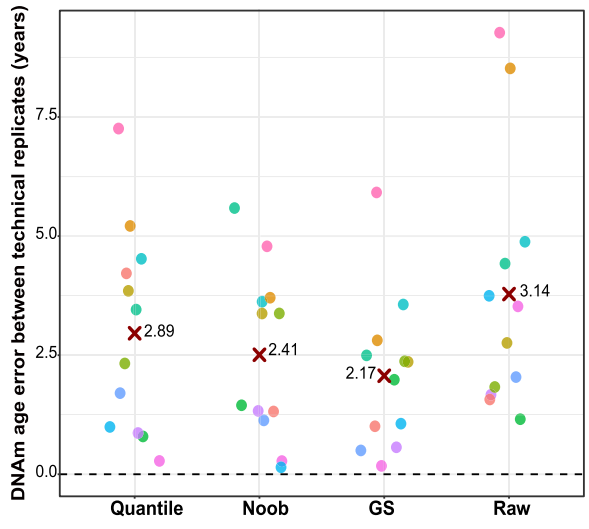

Fig. 4 Absolute difference between technical replicate pairs for each processing method. The *y* axis represents the absolute difference between each technical replicate pair's DNA methylation (DNAm) age. Each processing method is represented on the *x* axis. The median difference is indicated by the red cross for each group. The color of data points represents one technical replicate pair for ease of interpretation across methods. Error refers to the median absolute difference between DNAm age and chronological age

Finding that preprocessing method influenced mean values of DNAm age is important for the interpretation of future analyses, as we demonstrated that variation in DNAm age can be introduced by how the data are preprocessed. Our work here provides supporting evidence for the DNAm age acceleration residual measure, since this value is reflective of inter-individual variability within a measured dataset, and is, therefore, more comparable across studies. In contrast, the DNAm age difference, the crude difference between estimated DNAm age and chronological age, can be reflective of global DNAm shifts due to preprocessing methods.

Whichever measure of DNAm age is used (acceleration difference, residual, or age itself), there is an additional consideration that small effect sizes should be interpreted with caution. To highlight this point, we calculated DNAm age for technical replicates from raw data and three different preprocessing methods. We found that while there was some variability in DNAm age across technical replicates, regardless of preprocessing methods, the observed median absolute error in DNAm age for each method (2.7–3.14 years) was lower than the reported error of the epigenetic clock (3.6 years) [17]. GS-preprocessed data produced the tightest replicates, followed by noob and then quantile normalization, and the consistency between replicates was lowest when DNAm age was calculated from raw data. These findings may suggest using preprocessed data rather than raw data, but overall we emphasize the importance of considering the technical error of the epigenetic clock and caution interpretation of changes of less than 3.6 years.

To examine the appropriateness of using EPIC data to calculate DNAm age for future research, we took advantage of a cohort with DNAm data on the same individuals on both the 450K and EPIC arrays. It is crucial to examine whether the epigenetic clock can continue to be used on EPIC data, as the 450K platform is no longer available. There was high consistency between 450K and EPIC DNAm age estimates, and the lack of 19 CpG sites did not significantly affect the prediction accuracy of the epigenetic clock. Probe-wise correlation coefficients of the 334 common clock CpG sites across the 450K and EPIC were lower than anticipated; however, previous reports have demonstrated that the majority of EPIC probes are not well correlated to those of the 450K and that this is most prevalent at invariable probes [21]. These observations highlight the robustness of the multi-CpG predictors assessed, as despite the low probe-wise correlations the correlation between the estimated ages was highly correlated. This result is consistent with classical test theory in that error for any given probe is random, and largely uncorrelated with the error of other probes, and therefore these random effects would become redundant in a composite index like epigenetic

age [26]. The consistency in estimated age lends support to the strength of this predictive model on the EPIC platform and will allow users to continue applying this bioinformatic tool to continue to calculate DNAm age.

To further examine the application of EPIC array data to predict DNAm age, we estimated DNAm age in an independent EPIC array cohort. We observed correlations between EPIC DNAm age and chronological age that were comparable to previous reports, specifically in PBMCs and BAL samples. The strong association in PBMCs is consistent with previous reports of DNAm age in PBMCs as generated from 450K data [13, 18]. We observed less consistency in the brush samples; however, this tissue was not included in the training data of the 353-CpG epigenetic clock and so performance may not be reflective of EPIC array but rather be a property of clock itself. This is reinforced by our experiment removing the 19 clock CpGs not present on the EPIC array from the 450K data, where we observed a nearly perfect correlation with the 353-CpG data, suggesting that the loss of the 19 clock CpG sites did not influence the accuracy of the epigenetic clock.

There are limitations to this study that should be taken into consideration when interpreting the results. The primary datasets we investigated when comparing EPIC versus 450K estimated DNAm age were from monocyte samples, and although we found that the lack of 19 CpGs did not affect the pan-tissue DNAm age estimate in this specific cell type, those 19 CpGs may be important to estimate age in other tissues. Their importance to other tissues remains to be explored. Additionally, the methods applied in the current study should not be generalized across all studies. For example, global normalization methods, such as quantile normalization, are not appropriate in all cases as interesting biological information can be removed in datasets with large variation across samples, such as cancer compared to normal or multiple-tissue projects. Instead, the use of these data transformation methods should be considered on a study-by-study basis [11]. Furthermore, while we are cognizant there are several other available preprocessing options, for the purposes of our exploration and presentation of these data, we only assessed three of the most common methods.

In summary, we have investigated and confirmed that two commonly used methods of DNAm-based age estimation, the 353 CpG Horvath model and the 71-CpG Hannum model, were not compromised when using the latest human DNAm microarray platform, the EPIC array, which is lacking 19/353 CpG and 6/71 CpG targets, respectively. We have also tested whether DNAm age estimates were influenced by the preprocessing stage; for example, whether raw data generated differing results than normalized data. We assessed raw data and three different preprocessed inputs (noob-, quantile-, and GS-normalized) and found age estimates were different, but less than that of the reported error of the model. Related, we finally also provided support for using the age acceleration residual metric rather than the age acceleration difference in studies applying the epigenetic clock. Our work will provide researchers the confidence to investigate DNAm age using the EPIC array, as well as encourage users to critically consider the technical error of the epigenetic clock when interpreting future findings.

Additional file

Additional file 1: Table S1. Cohort characteristics. **Table S2.** Comparisons of age acceleration metrics derived from different data source inputs. **Figure S1.** Correlation heat-map of 59 common polymorphic control probes for 172 common samples run on the EPIC and 450K methylation platforms. **Figure S2.** Correlations between chronological age, 450K DNA methylation age, and EPIC DNA methylation age estimates from each data input. **Figure S3.** Probe-wise correlations of the 334 common clock CpGs across the 450K and EPIC arrays illustrated lower beta range associated lower correlation across platforms for a given CpG. **Figure S4.** Hannum DNA methylation age estimates for 450K (71 CpGs) versus EPIC (65 CpGs) for each preprocessed data type. **Figure S5.** DNA methylation (DNAm) age from a reduced epigenetic age predictor (334 CpGs) compared to the full epigenetic age predictor (353 CpGs) using the same 450K dataset. **Figure S6.** Diesel Exhaust Study III EPIC DNA methylation beta-value distribution across 795,882 sites. **Figure S7.** Heat map of Kendall rank coefficients across preprocessing methods in EPIC data. **Figure S8.** Dendrogram of 59 single nucleotide polymorphic control probes of technical replicates from the EPIC array. (PDF 1168 kb)

Abbreviations

27 K: Infinium Methylation27 BeadChip; 450K: Infinium Methylation450 BeadChip; BAL: Bronchoalveolar lavage; Brush: Bronchial brushing; CpG: Cytosine-phosphate-guanine; DE3: Diesel Exhaust III Study; DNAm: DNA methylation; EPIC: Infinium MethylationEPIC BeadChip; GS: GenomeStudio; Noob: Normal-exponential out-of-band; PBMC: Peripheral blood mononuclear cell; SNP: Single-nucleotide polymorphism

Funding

LMM is supported by a CIHR Frederick Banting and Charles Best Doctoral Research Award (F15-04283). CFR is supported by fellowships from the BC Lung Association, MITACS, and the Michael Smith Foundation for Health Research. CC is the Canada Research Chair in Occupational and Environmental Lung Disease. SH was supported by NIH/NIA U34AG051425-0. MSK is the Canada Research Chair in Social Epigenetics, Senior Fellow of the Canadian Institute for Advanced Research, and Sunny Hill BC Leadership Chair in Child Development. Support for the DE3 study was provided by AllerGen NCE, CIHR, and the BC Lung Association. Support for the EVOIMMUNOPOP study was funded by the Institut Pasteur, the CNRS and the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013)/ERC grant agreement 281297 (to LQ-M).

Availability of data and materials

The current study was a secondary use of the EVOIMMUNOPOP data and the authors of the original study [19] have made these DNAm data available at www.ncbi.nlm.nih.gov/geo/ under accession number GSE120610.

Authors' contributions

LMM designed and performed all data analyses and wrote the manuscript with MSK. RDE and MJJ contributed to the data analysis design and visualizations. JLM, DTSL, KE, and AM performed all DNAm arrays. LQ-M and LTH provided the EVOIMMUNOPOP DNAm cohort. CFR and CC contributed the DE3 Study DNAm data. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Approval for the monocyte EVOIMMUNOPOP project was approved by the Ethics Committee of the Ghent University, the Ethics Board of Institut Pasteur (EVOIMMUNPOP-281297) and the relevant French Authorities (CPP, CCITRS, and CNIL). Samples were collected after written informed consent had been obtained. Approval for the DE3 study was granted from the University of British Columbia Research Ethics Board (H11-01831).

Consent for publication
Not applicable.

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

Publisher's Note

Author details
[1]BC Children's Hospital Research Institute, Department of Medical Genetics, University of British Columbia, 950 West 28th Avenue, TRB A5-151, Vancouver, BC V5Z 4H4, Canada. [2]Department of Medicine, Division of Respiratory Medicine, University of British Columbia, Vancouver, BC, Canada. [3]Unit of Human Evolutionary Genetics, Institut Pasteur, 75015 Paris, France. [4]Centre National de la Recherche Scientifique (CNRS) UMR2000, 75015 Paris, France. [5]Center of Bioinformatics, Biostatistics and Integrative Biology, Institut Pasteur, 75015 Paris, France. [6]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA.

Received: 25 May 2018 Accepted: 1 October 2018
Published online: 16 October 2018

References
1. Aryee MJ, et al. Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30(10):1363–9.
2. Bibikova M, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1(1):177–200.
3. Bocklandt S, et al. Epigenetic predictor of age. PloS one. 2011;6(6):e14821.
4. Breitling LP, et al. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. Clin Epigenetics. 2016;8(1):21.
5. Davis S, Du P, Bilke S, Triche, Jr. T, Bootwalla M. methylumi: Handle Illumina methylation data. R package version 2.26.0; 2017.
6. Dedeurwaerder S, et al. A comprehensive overview of Infinium HumanMethylation450 data processing. Brief Bioinform. 2014;15(6):929–41.
7. Farré P, et al. Concordant and discordant DNA methylation signatures of aging in human blood and brain. Epigenetics Chromatin. 2015;8(1):19.
8. Fortin JP, Triche TJ Jr, Hansen KD. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics. 2016.
9. Garagnani P, et al. Methylation of ELOVL2 gene as a new epigenetic marker of age. Aging Cell. 2012;11(6):1132–4.
10. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359–67.
11. Hicks SC, Irizarry RA. Quantro: a data-driven approach to guide the choice of an appropriate normalization method. Genome Biol. 2015;16(1):117.
12. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115.
13. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. J Infect Dis. 2015;212(10):1563–73.
14. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat Rev Genetics. 2018;23:223.
15. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. Aging. 2015.
16. Horvath S, et al. Aging effects on DNA methylation modules in human brain and blood tissue. Genome Biol. 2012;13(10):R97.
17. Horvath S, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. Genome Biol. 2016;17(1):171.
18. Horvath S, et al. Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. Aging. 2015.
19. Husquin LT, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation. bioRxiv. 2018:371872.
20. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. Science (New York, NY). 2001;293(5532):1068–70.
21. Logue MW, et al. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. Epigenomics. 2017; 9(11):1363–71.
22. Marioni RE, Shah S, McRae AF, Chen BH, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015a;16(1):25.
23. Marioni RE, Shah S, McRae AF, Ritchie SJ, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian birth cohort 1936. Int J Epidemiol. 2015;44(4):1388–96.
24. Moran S, Arribas C, Esteller M. Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. 2015;8(3):389–99 dx.doi.org.
25. Morris TJ, et al. ChAMP: 450k Chip analysis methylation pipeline. Bioinformatics. 2014;30(3):428–30.
26. Novick MR, Lewis C. Coefficient alpha and the reliability of composite measurements. Psychometrika. 1967;32(1):1–13.
27. Schultz MD, et al. Human body epigenome maps reveal noncanonical DNA methylation variation. Nature. 2015;523(7559):212–6.
28. Teschendorff AE, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450k DNA methylation data. Bioinformatics. 2013;29(2):189–96.
29. Touleimat N, Tost J. Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics. 2012;4(3):325–41. https://doi.org/10.2217/epi.12.21.
30. Triche TJ, Daniel J. Weisenberger D, Van Den B, Laird PW, Kimberly D. Siegmund; Low-level processing of Illumina Infinium DNA Methylation BeadArrays, Nucleic Acids Research. 201;41(7):e90. https://doi.org/10.1093/nar/gkt090.
31. Triche TJ Jr, et al. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 2013;41(7):e90.
32. Troyanskaya O, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
33. Wang T, et al. A systematic study of normalization methods for Infinium 450K methylation data using whole-genome bisulfite sequencing data. Epigenetics. 2015;10(7):662–9.
34. Weidner CI, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014;15(2):R24.
35. Wilhelm-Benartzi CS, et al. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109(6):1394–402.
36. Yousefi P, et al. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. Epigenetics. 2014; 8(11):1141–52.