Factors Driving DNA Methylation Variation in Human Blood

Jacob Bergstedt[1,2,3,*], Sadoune Ait Kaci Azzou[1], Kristin Tsuo[1], Anthony Jaquaniello[1], Alejandra Urrutia[4], Maxime Rotival[1], David T. S. Lin[5], Julia L. MacIsaac[5], Michael S. Kobor[5], Matthew L. Albert[4], Darragh Duffy[6], Etienne Patin[1,8,*], Lluís Quintana-Murci[1,7-9,*], for the *Milieu Intérieur* Consortium

[1]Human Evolutionary Genetics, Institut Pasteur, UMR 2000, CNRS, Paris, France

[2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

[3]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[4]Insitro, San Francisco, USA

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

[6]Translational Immunology Laboratory, Institut Pasteur, Paris, France

[7]Chair of Human Genomics and Evolution, Collège de France, Paris, France

[8]These authors contributed equally

[9]Lead Contact

*Correspondence: jacob.bergstedt@ki.se (J.B.), epatin@pasteur.fr (E.P.), quintana@pasteur.fr (L.Q.-M.)

SUMMARY

Epigenetic changes are required for normal development and health, and can also underlie disease states; yet, the nature and respective contribution of factors that drive epigenetic variation in humans remain to be fully characterized. Here, we assessed how the blood DNA methylome of 958 adults is affected by genetic variation, aging, sex and 139 diverse environmental exposures, and investigated whether these effects are direct or mediated by changes in cellular composition, measured by deep immunophenotyping. We show that cellular heterogeneity and DNA sequence variation are the strongest predictors of DNA methylation levels. We identify latent cytomegalovirus infection as a major driver of DNA methylation variation and delineate three distinct effects of aging on DNA methylation, including increased dispersion consistent with epigenetic drift. Our rich dataset provides a unique resource for the design and interpretation of epigenetic studies and highlight critical factors in medical epigenomics studies.

Keywords:

epigenetics, DNA methylation, humans, blood, immune cells, aging, cytomegalovirus, smoking, sex, meQTLs, EWAS, gene-by-environment interactions, cellular deconvolution

INTRODUCTION

Epigenetic research has facilitated our understanding of the existing links between environmental risk factors, aging, genetic variation and human disease (Cavalli and Heard, 2019; Michalak et al., 2019). Epigenome-wide association studies (EWAS) have shown that DNA methylation (i.e., 5-methylcytosine, 5mC), the most studied epigenetic mark in human populations, is associated with a wide range of environmental exposures along the life course, such as chemicals, air pollution and nutrition (Martin and Fry, 2018), as well as past socioeconomic status (Bush et al., 2018; Karlsson Linner et al., 2017; Lam et al., 2012; Stringhini et al., 2015). Changes in DNA methylation have also been associated with non-communicable diseases, such as Parkinson's and Alzheimer's diseases, multiple sclerosis, systemic lupus erythematosus, type 2 diabetes and cardiovascular disease (Hwang et al., 2017; Ling and Ronn, 2019; Mazzone et al., 2019; van der Harst et al., 2017). These studies collectively suggest that DNA methylation marks could have tremendous value as a gauge of the exposome and as clinical biomarkers (Berdasco and Esteller, 2019; Wild, 2005).

However, interpretation of EWAS remains limited. First, because the epigenome of a cell reflects its identity (Farlik et al., 2016; Roadmap Epigenomics et al., 2015), a risk factor or a disease that alters cellular composition also alters 5mC levels measured in the tissue (Liu et al., 2013). It is thus necessary to determine if an exposure affects cellular composition or DNA methylation states of cell types, in order to better understand the link between such an exposure, DNA methylation and disease (Lappalainen and Greally, 2017). Previous studies have accounted for cellular heterogeneity in blood by using cell sorting experiments, or cellular proportions estimated from 5mC profiles through deconvolution techniques (Houseman et al., 2012; Teschendorff et al., 2017), but these approaches focus on a subset of frequent cell types that capture only a part of blood cellular composition. Second, genetic variation and DNA methylation are inextricably linked, as attested by the numerous DNA methylation quantitative trait loci (meQTLs) detected so far (Villicana and Bell, 2021), and genetic variants could confound associations between risk factors, 5mC levels and diseases or traits. Finally, environmental risk factors with a yet-unknown effect on DNA methylation, such as common infections, could also confound associations between other risk factors, DNA methylation and human phenotypes. Thus, a detailed study of the factors that impact DNA methylation at the population level, and the extent to which their effects are mediated by changes in cellular composition, is required to understand the role of epigenetic variation in health and disease.

To address this gap, we generated whole blood-derived DNA methylation profiles at >850,000 CpG sites for 958 healthy adults of the *Milieu Intérieur* cohort. We leveraged the deep characterization of the cohort, including high-resolution immunophenotyping by flow cytometry

(Patin et al., 2018; Thomas et al., 2015), to determine whether and how cellular composition, genetic variation, intrinsic factors (i.e., age and sex) and 139 diverse health-related variables and environmental exposures affect the blood DNA methylome. We performed EWAS adjusted or not for the measured proportions of 16 immune cell subsets, to robustly delineate effects on DNA methylation that are *direct*, i.e., acting through changes within cells, from those that are *mediated*, i.e., acting through subtle changes in cellular composition (Houseman et al., 2015). We find that the largest effects on DNA methylation are due to DNA sequence variation, whereas the most widespread differences among individuals are the result of blood cellular heterogeneity. We also identify latent cytomegalovirus (CMV) infection as a major driver of epigenetic variation and observe an increased dispersion of DNA methylation with aging, suggesting a decrease in the fidelity of the epigenetic maintenance machinery. Finally, we show that a large part of the effects on DNA methylation of aging, smoking, CMV serostatus and chronic low-grade inflammation is due to fine-grained changes in blood cell composition, and characterize the DNA methylation signature of cell-types affected by these factors. This work generates new hypotheses about mechanisms underlying DNA methylation variation in the human population and highlights critical factors to be considered in medical epigenomics studies.

RESULTS

Population Variation in DNA Methylation Differs Markedly Across the Genome

To investigate the contributions of genetic and non-genetic factors to population variation in DNA methylation, we quantified 5mC levels at >850,000 CpG sites, with the Illumina Infinium MethylationEPIC array, in the 1,000 healthy donors of the *Milieu Intérieur* cohort (Thomas et al., 2015). The cohort includes individuals of Western European origin, equally stratified by sex (i.e., 500 women and 500 men) and age (i.e., 200 individuals from each decade between 20 and 70 years of age), who were surveyed for detailed demographic and health-related information, including dietary habits, upbringing, socioeconomic status, mental health, past and latent infections, and vaccination and medical histories (Table S1). To adjust for and quantify the blood cell-composition-mediated effect on DNA methylation of genetic factors, intrinsic factors (i.e., age and sex) and environmental exposures, we measured, in all donors, the proportions of 16 major and minor immune cell subsets by standardized flow cytometry, including neutrophils, basophils, eosinophils, monocytes, natural killer (NK) cells, dendritic cells, B cells, $CD4^+$ and $CD8^+$ T cells at four differentiation stages (naive, central memory, effector memory and terminally differentiated effector memory cells) and $CD4^-CD8^-$ T cells (Patin et al., 2018). Furthermore, all donors were genotyped at 945,213 single-nucleotide polymorphisms (SNPs), yielding 5,699,237 accurate SNPs after imputation (Patin et al., 2018). After quality control filtering, high-quality measurements of 5mC levels were obtained at 644,517 CpG sites for 958 unrelated individuals (Figure S1; **STAR Methods**).

We first investigated population variation in DNA methylation across different well-characterized chromatin states, using naïve $CD4^+$ T cells as a reference (Roadmap Epigenomics et al., 2015). We found that CpG sites in transcription start sites (TSS) are typically unmethylated and exhibit the lowest population variance in 5mC levels (Figure 1A, B), suggesting that epigenetic constraints are the strongest in promoters, whereas actively transcribed gene bodies and hetero-chromatin are highly methylated and also show low population variance. 5mC levels in enhancers and Polycomb-repressed regions are the most variable (Figure 1A, B), suggesting that DNA meth-ylation in these regions are preferentially affected by genetic, intrinsic or environmental factors, or cellular heterogeneity. These results indicate that 5mC measurements from our cohort reproduce the known properties of DNA methylation and show high levels of variation across the epigenome and among individuals.

Widespread Local Genetic Control of DNA Methylation in Whole Blood

Studies of DNA methylation quantitative trait loci (meQTLs) have revealed that DNA sequence variation affects 5mC levels at numerous nearby CpG sites (Bonder et al., 2017; Hannon et al., 2018), a feature that can confound associations between environmental exposures and DNA methylation (Lappalainen and Greally, 2017). To account for this, we estimated, for each CpG site, the effects of nearby DNA sequence variation on their 5mC levels (100-Kb window; **STAR Methods**). We adjusted models on genetic ancestry and blood cell-type proportions, as well as age, sex, smoking status and CMV serostatus, which we have previously shown to affect blood cell composition (Patin et al., 2018). We found a significant local meQTL for 29.2% of the 644,517 tested CpG sites (n = 188,129; two-stage adjusted P-value, $P_{adj} < 0.05$; Figure S2; **STAR methods**). We detected 1,978 CpG sites with a difference in the proportion of DNA methylation between homozygotes larger than 0.3, indicating that the effect of local meQTLs can be substantial.

We found that CpG sites with a local meQTL are enriched in enhancers (odds ratio [OR] 95% CI: [1.87, 1.95]; Figure 1C; **STAR Methods**), where the population variance of DNA methylation is the largest (Figure 1B). Conversely, CpG sites with a local meQTL are depleted in TSS and actively transcribed genes (OR CIs: [0.32, 0.34] and [0.49, 0.52]), where DNA methylation variance is the lowest and sequence conservation the highest (Figure 1D). While confirming that local meQTLs are enriched in disease and trait associations by genome-wide association studies (GWAS) (Bonder et al., 2017), we found that the enrichment is strongest in enhancers and genic enhancers (enrichment > 1.8; $P_{resampling} < 1.0\times10^{-4}$; Figure 1E) and absent from regions of low sequence conservation, such as heterochromatin. These findings indicate that DNA sequence variants have widespread, strong effects on nearby DNA methylation levels, particularly in regulatory elements, and that local meQTLs are enriched in genetic variants that affect phenotypic variation and disease risk.

Structural Factors and Zinc Finger Proteins are Regulators of DNA Methylation

We investigated the long-range genetic control of DNA methylation, by estimating the effect of genome-wide variants on 5mC levels of a selection of CpG sites, to reduce the burden of multiple testing. We selected the 50,000 CpG sites with the highest residual variance after fitting a model including as predictors: (i) the most associated local meQTL variant, (ii) genetic ancestry, (iii) blood cell proportions and (iv) non-genetic factors affecting blood cell composition (**STAR Methods**). We found 2,394 independent long-range meQTLs, for 1,816 CpG sites (3.6%) and 1,761 independent variants (Table S2). The effects of long-range meQTLs are generally weaker than those of local meQTLs (Figure S2), yet we found 152 CpG sites with a difference in the proportion of methylation between homozygotes larger than 0.15. As for local meQTLs, remote-effect variants are also

enriched in GWAS hits (enrichment = 1.78; $P_{resampling} < 6.3\times10^{-5}$). Interestingly, CpG sites under remote genetic control are enriched in TSS regions and regions associated with *ZNF* genes, in contrast with CpG sites under local genetic control (Figure 1C, F). Furthermore, we found that remote meQTL SNPs are strongly concentrated in *ZNF* genes (OR CI: [15.1, 25.8]; Figure 1G). These findings indicate that zinc-finger proteins (ZFPs) play a role in the long-range control of DNA methylation, in line with their role in the regulation of heterochromatin and *ZNF* gene expression (Marchal and Miotto, 2015; O'Geen et al., 2007; Quenneville et al., 2012).

Genetic variants controlling the expression of nearby transcription factors (TF) have been found to have widespread, long-range effects on the DNA methylome (Bonder et al., 2017; Hop et al., 2020). Of the 3,643 genes with a nearby variant associated with a remote CpG site, 33% have its expression altered by the same variant in the eQTLGen database (Võsa et al., 2018). In total, 200 protein-coding genes have local variants that are associated with 5mC levels at ≥ 10 remote CpG sites (Figure 1H). Variants that affect the largest number of remote CpG sites are located nearby well-known structural factors and TFs such as *SENP7*, *BCLAF1*, *CTCF*, *NFKB1* and *NFE2*, and, consistently, CpG sites remotely associated with a TF local eQTL are strongly enriched in binding sites of the corresponding TF, or a TF related to it (Tables 1 and S3). For example, the rs10417143 variant alters *ZNF257* mRNA levels ($P_{adj} = 2.8\times10^{-70}$; (Võsa et al., 2018)) and 5mC levels at 16 CpG sites, which are enriched in binding sites for the ZNF534 TF (Figure 1I). Likewise, rs12491955 is associated with increased *SENP7* mRNA levels ($P_{adj} = 4.2\times10^{-302}$; (Võsa et al., 2018)) and 5mC levels at 35 CpG sites, 30 of which are located in two clusters of *ZNF* genes on chromosome 19 (Lemire et al., 2015). Of these 35 CpG sites, 23 are located in binding sites for KAP1 (encoded by *TRIM28*, Figure 1J), a chromatin remodeler regulated by SENP7 (Garvin et al., 2013). Of note, 20 out of the 50 most enriched TFs are ZFPs (Table S3). These results collectively support the notion that transcriptional variation of TFs results in DNA methylome-wide changes due to differential occupancy of their binding sites and highlight the role of chromatin remodelers and ZFPs in the regulation of DNA methylation.

Aging Elicits DNA Hypermethylation Related to Polycomb Repressive Complexes and Increased Epigenetic Dispersion

The link between DNA methylation and aging is well established (Hannum et al., 2013; Heyn et al., 2012; Johansson et al., 2013; Jones et al., 2015; Wang et al., 2018); however, given that blood cell composition is altered with age (Patin et al., 2018), it remains unclear how age impacts DNA methylation in a heterogeneous tissue such as blood (Jaffe and Irizarry, 2014). We thus investigated how the DNA methylome is shaped by the intertwined processes of cellular aging (i.e., direct, cell-

composition-independent effects of age) and age-related changes in blood cellular composition (i.e., effects of age mediated by changes in cellular composition) (**STAR Methods**). At a false discovery rate (FDR) of 0.05, we found that age has a significant total (i.e., direct and mediated) effect on 5mC levels at 258,830 CpG sites (40.2% of CpG sites; Figure 2A and Table 2). When estimating direct effects by adjusting on measured immune cell proportions, a significant effect was observed at 144,114 CpG sites (22.4% of CpG sites; FDR < 0.05; Table 2), highlighting the widespread, cell-composition-independent effect of age on 5mC levels in blood. Importantly, when we applied a commonly used deconvolution method to correct for cellular heterogeneity (Koestler et al., 2016), the number of age-associated CpG sites almost doubled (n = 277,209), suggesting that corrections based on estimated proportions of major cell subsets are incomplete.

We observed that 69% of CpG sites directly associated with age show a decrease in 5mC levels (Figure 2B, C). This pattern predominates in quiescent chromatin, actively transcribed genic regions and enhancers. In contrast, DNA hypermethylation was observed in 87% of age-associated CpGs within CpG islands (CGIs; Figure S3A, B). Consistently, CpG sites exhibiting increasing 5mC levels with age are predominantly found in Polycomb-repressed regions, bivalent TSSs and bivalent enhancers, which are CGI-rich regions (Figures 2B, C and S1K, L). Furthermore, these CpG sites are the most enriched in binding sites for RING1B, JARID2, RYBP, PCGF1, PCGF2 and SUZ12 TFs (OR > 10.0; Figure 2D and Table S3), which are all part of the Polycomb repressive complexes (PRC) 1 and 2. PRC1 and PRC2 mediate cellular senescence and modulate longevity in invertebrates (Bracken et al., 2007; Siebold et al., 2010). Importantly, when we restricted the analysis to CpG sites outside of CpG islands, we found similar enrichments in Polycomb-repressed regions (OR 95% CI [10.4, 14.0]) and PRC TF binding sites (RING1B OR 95% CI: [11.4, 14.1]; PCGF2 OR 95% CI [9.89, 13.5]). Finally, genes with age-increasing methylation levels are strongly enriched in developmental genes ($P_{adj} = 2.2\times10^{-42}$; Table S4), which are regulated by PRCs (Boyer et al., 2006). These results support a key regulatory role of Polycomb proteins in age-related hypermethylation (Dozmorov, 2015).

Global hypomethylation of the genome and CGI-associated hypermethylation are both hallmarks of cancer (Timp and Feinberg, 2013). We found that genes with a TSS that is increasingly methylated with age are significantly enriched in tumor suppressor genes (OR = 1.55, CI: [1.28, 1.85]; Fisher's exact test $P = 4.0\times10^{-6}$) (Zhao et al., 2016). For example, 5mC levels linearly increase by 0.2% per year of age nearby the TSS of *BCL6B* ($P_{adj} = 4.6\times10^{-66}$; Figure 2E), a tumor suppressor gene that is hypermethylated in cancer (Xu et al., 2012a). In addition, 5mC levels increase by 0.1% per year of age nearby the TSS of *DNMT3A* ($P_{adj} = 6.6\times10^{-17}$; Figure S3E), which encodes a DNA methyltransferase that plays a key role in tumorigenesis (Jost et al., 2014). Our

results indicate that genomic hypomethylation and CGI-associated hypermethylation are lifelong progressive processes, possibly due to an altered maintenance of DNA methylation after multiple mitotic cell divisions (Teschendorff et al., 2010; Yang et al., 2016; Zhou et al., 2018), and support an intricate link between aging and oncogenesis.

Finally, we determined whether the variance of 5mC levels among individuals depends on age, a phenomenon known as "epigenetic drift" (i.e., the divergence of the DNA methylome as a function of age due to stochastic changes) (Fraga et al., 2005; Jones et al., 2015), by fitting models parameterizing the residual variance with a linear age term, and adjusting for cellular composition, aging, CMV serostatus and sex in the mean function (**STAR methods**). We observed a significant dispersion with age for 16.3% of all CpG sites. Strikingly, 90% of these CpGs show an increase in the variance of 5mC levels with age (Figure 2F), suggesting a decrease in the fidelity of epigenetic maintenance associated with aging. Examples of CpG sites with large, age-increasing dispersion are found in the TSS of the *MAFA* and *CBLN1* genes ($P_{adj} = 2.08\times10^{-49}$ and 4.63×10^{-45}; Figure 2G, H). Similar results were obtained when adjusting the variance function for cellular composition (**STAR methods**). In addition, we found that, out of 104,786 CpG sites with age-related dispersion, 63% show no significant changes in 5mC levels with age (Figure S3F), which suggests that these results are not driven by relationships between the average and variance of 5mC levels. Collectively, these findings indicate that aging elicits numerous DNA methylation changes in a cell-composition-independent manner, including global epigenome-wide demethylation, hypermethylation of PRC-associated regions, and increased variance, highlighting the occurrence of different mechanisms involved in epigenetic aging.

Immunosenescence-Related Changes in Cellular Composition Mediate DNA Methylation Variation with Age

We detected a significant cell-composition-mediated effect of age at ~12% of CpG sites (n = 75,301; Table 2), indicating that a substantial fraction of age-associated changes in DNA methylation are due to age-related changes in immune cell proportions. However, mediated effects are typically weaker than direct effects (Figure 2C, K) and CpG sites with the strongest direct age effects show no mediated effect, suggesting that their changes in 5mC levels are shared across cell types (Figure 2A). In contrast to direct effects, mediated effects are most often associated with demethylation, regardless of their localization (Figures 2K and S3C, D). Yet, enhancers and TSS-flanking regions (but not TSS themselves) are enriched in CpG sites with a significant cell-composition-mediated, positive effect of age (Figure 2J), possibly because these regions tend to be active in a cell-type specific manner (Roadmap Epigenomics et al., 2015). In addition, CpG sites

with a cell-composition-mediated increase in DNA methylation are enriched in TF binding sites for RUNX1 and RUNX3, two key regulators of hematopoiesis (Figure 2L and Table S3). Genes with CpG sites showing a mediated increase or decrease in DNA methylation with age are respectively enriched in genes involved in lymphoid or myeloid cell activation ($P_{adj} < 2.0\times10^{-14}$; Table S4), indicating that mediated age effects on DNA methylation are related to progressive differences in the composition of the lymphoid and myeloid compartments.

We then determined if age elicits 5mC changes specific to certain immune cell subsets, by deriving and verifying an interaction model capable of dealing with the compositional nature of immune cell proportions in blood, similarly to previous work (Zheng et al., 2018) (**STAR Methods**). Because inference was unstable for rare cell subsets, we restricted the analysis to six major immune cell types ($CD4^+$ and $CD8^+$ T cells, $CD4^-$ $CD8^-$ T cells, B cells, NK cells, and monocytes), and estimated how DNA methylation changes with age within these cell types, compared to the most frequent cell type, i.e., neutrophils. We found that 17% of tested CpG sites (n = 106,899) show a cell-type-dependent association with age in $CD8^+$ T cells (Figure 2I), 69% of which show decreasing 5mC levels (Figure S3G). These results support previous studies reporting that this T cell subset undergoes substantial, lifelong epigenetic changes (Goronzy et al., 2018; Tserel et al., 2015). Together, our findings provide strong statistical evidence that DNA methylation variation with age rely on different, non-mutually exclusive mechanisms: the progressive decline of the epigenetic maintenance system, common to all cell types, and the increased heterogeneity of immune cell subsets that characterizes immunosenescence (Nikolich-Zugich, 2018).

Sex Differences in DNA Methylation are Predominantly Cell- and Age-Independent

Given that substantial differences in immune cell composition between sexes have been observed (Patin et al., 2018), we next assessed how cellular heterogeneity contributes to sex differences in DNA methylation (Singmann et al., 2015; Yousefi et al., 2015). We found ~29% of CpG sites (n = 186,545) with a significant total effect of sex, ~20% (n = 126,904) with a significant direct effect, and ~7% (n = 44,667) with a significant cell-composition-mediated effect (FDR < 0.05; Table 2 and Figure S4A). The largest direct effects of sex were observed at *DYRK2*, *DNM1*, *RFTN1*, *HYDIN*, and *NAB1* genes ($P_{adj} < 1.0\times10^{-285}$). For example, the *DYRK2* promoter is 11% and 45% methylated in men and women, respectively, at the CpG site bound by the X-linked PHF8 histone demethylase (Figure S4B, C). DYRK2 phosphorylates amino acids and plays a key role in breast and ovarian cancer development (Correa-Saez et al., 2020).

DNA methylation levels were higher in women at 78% of sex-associated autosomal CpG sites (Figure S4D, E), a pattern also observed in newborns (Yousefi et al., 2015). This proportion is

similar across different genomic regions, based on either chromatin states or CpG density (Figure S4E, I). When quantifying how sex differences in DNA methylation vary during adulthood, by adding a sex-by-age interaction term to our models (**STAR Methods**), we found only 23 CpG sites with a significant, sex-dependent effect of age (FDR < 0.05; Table S5), confirming previous findings (McCartney et al., 2019; Yusipov et al., 2020). The most associated genes are *FIGN*, associated with risk-taking behaviors (Karlsson Linner et al., 2019) and educational attainment (Lee et al., 2018), and *PRR4*, associated with the dry eye syndrome, a hormone-dependent, late-onset disorder (Perumal et al., 2016). Overall, our findings indicate that the blood DNA methylome is widely affected by sex, but its effects are typically not mediated by cellular composition and do not change during adulthood.

Cytomegalovirus Infection Alters the Blood DNA Methylome through Regulation of Host Transcription Factors

We next leveraged the extensive questionnaire and phenotyping conducted in the *Milieu Intérieur* cohort to identify environmental factors that elicit cell-composition-independent changes in the blood methylome. Specifically, we estimated how 5mC levels are influenced by 139 variables (Table S1), including factors related to upbringing (e.g., birth weight, delivery route, rural or urban childhood), socio-economic status (e.g., educational attainment, monthly income, work-hours), dietary habits (e.g., eating frequency of various foods), health-related habits (e.g., smoking, BMI, physical exercise), lipid metabolism (e.g. low- and high-density lipoproteins, cholesterol and triglyceride levels), mental health and sleeping habits (e.g., self-reported depression, hours of sleep), exposure to pollutants (e.g., asbestos, benzene, silica), reproductive life cycle and contraception in women (e.g., contraceptive use, age at menopause), past and present exposure to infectious agents (e.g., cytomegalovirus, Epstein-Barr virus), total serum antibody concentrations (e.g. levels of IgG, IgE and IgM), and vaccination history (e.g., MMR, hepatitis A vaccine). Tests between each of the variables and 5mC levels at all measured CpG sites were considered a separate family and were adjusted to control the FDR at 0.05 (**STAR Methods**). All models were adjusted for associated meQTLs, genetic ancestry, batch variables and factors that impact cell composition, including sex and a non-linear age term.

The factor that is associated with the largest number of CpG sites is CMV serostatus. CMV is the causative agent of a latent, mainly asymptomatic, infection with prevalence ranging from 40% to 100% (Cannon et al., 2010), which drastically alters the composition of the $CD8^+$ T cell compartment in blood (Klenerman and Oxenius, 2016). CMV seropositivity has a significant total effect on ~36% of CpG sites (n = 233,014; Figure 3A and Table 2). When adjusting for blood cell

composition, a significant direct effect was detected for ~10% of CpG sites (n = 64,383; FDR < 0.05). Of note, the 16 cell proportions we adjusted for include central, effector memory and EMRA $CD8^+$ T cells, which we have previously shown to be strongly associated with CMV serostatus (Patin et al., 2018). When we used the standard deconvolution method to correct for cellular heterogeneity, which does not include estimates of $CD8^+$ sub-compartments (Koestler et al., 2016), we found twice as many CpG sites directly associated with CMV serostatus (~19%, n = 120,024), indicating again that this standard correction for cellular heterogeneity is not complete.

One of the strongest direct effects of CMV infection was observed nearby the TSS of *LTBP3* (β value scale 95% CI: [3.0%, 4.2%], $P_{adj} = 2.2\times10^{-33}$; Figure 3E). LTBP3 is a regulator of latent transforming growth factor β (TGF-β) (Morita et al., 2016), which is induced in CMV latently infected cells (Mason et al., 2012). We found that effects of CMV are typically smaller than those of age and sex (Figure S2) and are associated with an increase in 5mC levels in 92% of CpG sites within CGIs and a decrease in 76% of CpGs outside CGIs (Figure S5A, B). As for age, we observed in CMV^+ donors an overall increase in 5mC levels in Polycomb-repressed regions and binding sites of PRC-related TFs, and a decrease in regions of strong transcription (Figure 3B-D), suggesting dysregulation of the host gene inactivation machinery as a result of latent infection. Similar results were found when restricting the analysis to CpG sites outside of CGIs (Polycomb-repressed regions OR = 2.9, CI: [1.9, 4.24]). Interestingly, CpG sites showing increased 5mC levels in CMV^+ donors are strongly enriched in binding sites for the BRD4 TF (OR = 13.5, CI: [11.9, 15.3], $P_{adj} < 1.0\times10^{-320}$; Figure 3D and Table S3), a bromodomain protein that plays a critical role in the regulation of latent and lytic phases of CMV infection (Groves et al., 2021). In addition, CpG sites showing a decrease in DNA methylation in CMV^+ donors are strongly enriched in binding sites for BATF3 (OR = 9.0, CI: [8.4, 9.7], $P_{adj} < 1.0\times10^{-320}$; Figure 3F and Table S3), which is paramount in the priming of CMV-specific $CD8^+$ T cells by cross-presenting dendritic cells (Torti et al., 2011).

We investigated if the large shift in the composition of the $CD8^+$ T cell compartment caused by CMV (Klenerman and Oxenius, 2016) is accompanied by changes in 5mC levels. We found that 33% of CpG sites show a significant cell-composition-mediated effect of CMV serostatus (n = 217,223, FDR < 0.05). Importantly, 93% of CpG sites with a significant direct effect also show a significant mediated effect (n = 60,194; Figure 3A), and we observed a clear correlation between total and mediation effect sizes (R = 0.93; Figure 3A). For example, the CpG site in the TSS of *LTBP3* with a large direct effect of CMV has also a large, mediated effect (CI: [2.4%, 3.5%], $P_{adj} = 1.8\times10^{-18}$); CMV^+ donors show higher proportions of $CD8^+$ T_{EMRA} cells ($P = 1.38\times10^{-49}$), which in turn are associated with higher 5mC levels at *LTBP3* ($P = 6.9\times10^{-92}$), supporting mediation by this T cell subset (Figure 3G, H). Collectively, our analyses indicate that CMV infection affects a large

fraction of the human blood DNA methylome through the dysregulation of host TFs and fine-grained changes in cellular composition.

Strong Effects of Smoking are Reversible and Independent of Blood Cell Composition

The second exposure that is associated with the largest number of CpG sites is cigarette smoking, for which the total effect was significant for 7,257 CpG sites (~1.1%, Figure 3I and Table 2). Although active smoking is known to elicit reproducible changes in DNA methylation (Dugue et al., 2020; Gao et al., 2015), we previously showed that smoking also has a broad effect on blood immune cell subsets (Patin et al., 2018), suggesting possible mediation by cellular composition. When adjusting for the 16 immune cell proportions, we found that smoking directly alters 5mC levels at 2,416 CpG sites (~0.4% of CpG sites; FDR < 0.05; Table 2), 62% of which show a decrease in 5mC levels. For example, smokers show strongly decreased 5mC levels in the introns of the dioxin receptor repressor gene *AHRR* (β value scale 95% CI: [-23%, -20%], $P_{adj} = 9.9\times10^{-128}$), the second exon of *F2RL3* (CI: [-10%, -8.6%], $P_{adj} = 3.2\times10^{-78}$) and the first intron of *RARA* (CI: [-10%, -8.5%], $P_{adj} = 5.4\times10^{-67}$), in agreement with previous studies (Dugue et al., 2020; Gao et al., 2015). No clear differences in the distribution of effect sizes were observed over genomic regions (Figure S5E, F). CpG sites that are demethylated in smokers are significantly enriched in binding sites for the hypoxia-related TFs EPAS1 and HIF2A ($P_{adj} < 7.0\times10^{-6}$), as well as AHRR (OR = 7.22, CI: [3.91, 12.2], $P_{adj} = 2.4\times10^{-8}$; Figure 3J and Table S3). This indicates that AHRR up-regulation in smokers elicits decreased 5mC levels at AHRR binding sites. To determine if such direct effects are reversible, we compared, for all smoking-associated CpG sites, the changes in 5mC levels with years since last smoke for past smokers, to the changes with years of smoking for active smokers. In agreement with previous studies (Dugue et al., 2020; Gao et al., 2015), we found a strong negative correlation between effect sizes (R = -0.70; slope = -1.12; Figure S5I), supporting the reversibility of the direct effects of smoking on DNA methylation.

We estimated that 5mC levels are significantly altered by smoking due to changes in cell composition at ~3.2% of CpG sites (n = 20,381 CpG sites with a mediated effect; FDR < 0.05). Among the most strongly affected CpG sites, we found *IL18RAP* (CI: [0.85%, 1.48%], $P_{adj} = 1.23\times10^{-9}$), a subunit of the receptor for IL18 that is differentially expressed by NK cells (Crinier et al., 2018). We observed that active smoking induces a reduction in the proportion of NK cells ($P = 1.17\times10^{-10}$), which is in turn associated with lower 5mC levels at *IL18RAP* ($P = 2.77\times10^{-29}$; Figure 3K, L), in line with an effect of smoking mediated by NK cells. Of note, mediated effects of smoking on 5mC levels were also reversible, to a degree similar to that of direct effects (R = -0.69; slope=-0.84, Figure S5J). Importantly, only 5.3% of CpG sites with a direct effect of smoking status

also have a significant mediation effect, and, on average, mediated effects of smoking are weaker than direct effects (Figure 3I). Out of the 50 CpG sites with the largest total effect of smoking status ($P_{adj} < 2.4\times10^{-18}$), 49 CpG sites, including those in *AHRR*, showed no significant mediation effect (FDR < 0.05). Collectively, these findings indicate that the largest effects of cigarette smoking on the blood DNA methylome are reversible and independent of blood cell composition.

Other Environmental Exposures do not Trigger Strong, Widespread Changes in the Adult DNA Methylome

The third environmental exposure that we identified as affecting DNA methylation variation is circulating levels of C-reactive protein (CRP), a marker of chronic, low-grade inflammation in healthy adults. Associations between CRP levels and hundreds of 5mC marks have been detected (Ligthart et al., 2016), but the strong relationship of CRP levels with the immune system (Sproston and Ashworth, 2018) and genetic variation (Ligthart et al., 2018) suggests that that these factors could confound associations. Specifically, changes in blood cell composition may be the cause of changes in CRP levels, and this could induce spurious associations, instead of mediated effects, at CpG sites associated with immune cell proportions.

We found an association between CRP and 5mC levels at 20,043 CpG sites (~3.1% of CpG sites; FDR < 0.05; Table 2), a figure that, when adjusting for cellular composition, dropped to only 480, of which 80% (n = 386) showed decreased 5mC levels with increased CRP levels. We detected a CpG site within an enhancer nearby *BCL2*, a key regulator of apoptosis and inflammation (Chong et al., 2020), where 5mC levels increase with increasing CRP levels (β value scale 95% CI: [0.6%, 1%], $P_{adj} = 1.06\times10^{-5}$; Figure S5K). Another example is a CpG site within an enhancer nearby *ABCG1* (CI: [0.4%, 0.8%], $P_{adj} = 1.20\times10^{-5}$; Figure S5L). In our cohort, 5mC levels at the same site are also associated to triglyceride (CI: [1.2%, 2.1%], $P_{adj} = 2.42\times10^{-8}$) and HDL (CI: [-3.9, -2.1], $P_{adj} = 9.35\times10^{-9}$) levels. The associations were retained in a model including CRP, HDL and triglyceride levels, indicating that they affect *ABCG1* 5mC levels independently. CRP is known to inhibit cellular cholesterol efflux by downregulating *ABCG1* mRNA levels, which are impaired in patients with type 2 diabetes, obesity, and hypertension (Li et al., 2012). These results indicate that associations between CRP levels and DNA methylation are mainly, but not exclusively, due to changes in blood cell composition, and generate new hypotheses on the epigenetic mechanisms relating subclinical inflammation to metabolic conditions.

Besides CMV infection, smoking status and chronic inflammation, we found limited evidence of a direct effect of non-heritable factors on the DNA methylome of healthy adults (Table 2). We found a significant total effect of heart rate, ear temperature and hour of blood draw on 5mC levels at

76,018, 59,728 and 38,884 CpG sites, respectively (FDR < 0.05), but no associations remained significant when adjusting for cellular heterogeneity. In total, we found 59 significant cell-composition-independent associations between the remaining non-heritable factors and 5mC levels (Table 2), the majority of which relate *ABCG1*, *DHCR24* and *CPT1A* genes with lipid-related traits and BMI (Braun et al., 2017). In addition, we found a single association between log protein levels and 5mC levels at a CpG site close to *DAO*, a gene encoding D-amino acid oxidase involved in protein catabolism (95% CI: [3.4%, 7%], $P_{adj} = 0.015$). We detected an association close to *TCERG1L* and educational attainment (95% CI: [-0.028%, -0.013%], $P_{adj} = 0.019$). Genetic variation in *TCERG1L* is associated with years of education (Lee et al., 2018), but not with 5mC levels in our cohort. We also found a significant association between log uric acid levels and 5mC levels at the *SLC2A9* gene (cg00071950; $P_{adj} = 0.0034$), which is no longer significant when adjusting on the local meQTL SNP ($P_{adj} = 1.0$), illustrating how DNA sequence variation can confound EWAS results. Nutritional habits, assessed based on 20 dietary frequency variables, have no detectable effects on the blood DNA methylome, except the frequency of raw fruit consumption at *GLI2* (95% CI: [-2.5%, -1.1%], $P_{adj} = 0.0022$). Of note, we did not replicate previously reported associations between DNA methylation and serum IgE levels (Ek et al., 2017; Liang et al., 2015) and did not detect any association with current socio-economic status. Collectively, these results indicate that environmental exposures related to upbringing, socio-economic status, nutrition or vaccination do not induce strong changes of the blood DNA methylome in our cohort of healthy adults.

Gene × Environment and Gene × Cell Type Interactions Affect DNA Methylation Variation

Gene × environment interactions are thought to underlie adaptable human responses to environmental exposures through epigenetic changes (Feinberg, 2018). Having established that age, sex, CMV serostatus, smoking status and chronic low-grade inflammation (CRP levels) are the main non-heritable determinants of DNA methylation variation, we evaluated whether their effects are genotype-dependent. We thus tested for genotype × age, genotype × sex or genotype × exposure interactions, adjusting for 16 measured cell proportions (**STAR Methods**). We found evidence of genotype-dependent effects at 175, 41, 4, 29 and 0 CpG sites for age, sex, smoking status, CMV serostatus and CRP levels, respectively ($P_{adj} < 0.05$, MAF > 0.10; Figure 4A; Table S5), the interacting SNP being local in all except 7 cases. We detected a strong genotype × age interaction for three CpG sites located in the *BACE2* gene, the 5mC levels of which decrease with age only in donors carrying the nearby rs2837990 G>A allele (β value scale 95% CI: [11%, 13%], $P_{adj} = 2.83\times10^{-10}$; Figure 4B; Table S5). *BACE2* encodes beta-secretase 2, one of two proteases involved

in the generation of amyloid beta peptide, a critical component in the etiology of Alzheimer's disease (Holler et al., 2012).

We then explored whether genetic variants affect 5mC levels specifically in different immune cell types, i.e., cell-type-dependent meQTLs. Because inferences were unstable for rare immune cell subsets, we estimated the effects of associated variants within six major cell types, compared to the effect of the variants within neutrophils – the most frequent blood cell subset. We found that genotypes affected DNA methylation differently according to cellular composition at 695 CpG sites. We found 264, 157, 62, 56, 32, and 19 significant interaction effects for $CD4^+$ T cells, $CD8^+$ T cells, NK cells, B cells, $CD4^-CD8^-$ T cells and monocytes, respectively ($P_{adj} < 0.05$; Figure 4A). One of the strongest signals was found between 5mC levels at the TSS of *CD300A* and the nearby rs12939435 variant, the effects of which depend on the proportion of $CD8^+$ T cells (CI: [-0.50%, -0.29%], $P_{adj} = 2.19\times10^{-14}$; Figure 4C; Table S5). CD300A is an immunomodulatory molecule that is expressed in various immune cell types and is associated with a cytotoxic molecular signature in $CD8^+$ T cells (Xu et al., 2012b). Overall, our analyses identify several environment- and cell-type-dependent meQTLs, supporting a strong, but limited impact of gene × environment and gene × cell type interactions on the blood DNA methylome.

Genetics and Cellular Heterogeneity Drive DNA Methylation Variation in Human Blood

Having established how genetic variation, cellular composition, intrinsic factors and a broad selection of non-heritable factors shape the blood DNA methylome, we next sought to compare the relative impact of these factors on DNA methylation. We classified the factors into four groups: (i) the cellular heterogeneity group, which consists of the 16 measured cell proportions; (ii) the intrinsic group, which consists of age and sex; (iii) the genetic group, which consists of the most associated local-meQTL variant around each CpG site; and (iv) the exposure group, which consists of smoking status, CMV serostatus and chronic low-grade inflammation. Since these groups vary in their degrees of freedom, we measured the relative predictive strength for each CpG site by the out-of-sample prediction accuracy, estimated by cross-validation (**STAR methods**). To ensure unbiased estimates, we mapped local meQTLs anew within each training set.

The model explains < 5% of out-of-sample variance for 51% of CpG sites (Figure 5A), which are typically characterized by low total 5mC variance (Figure S6A). This suggests that these sites are constrained in the population and that small fluctuations in 5mC levels determine their variation, possibly due to measurement errors or biological noise. Nevertheless, the model explains > 25% of DNA methylation variance for 21% of CpG sites (n = 133,180). The strongest predictor for these CpGs is cellular composition, genetics, intrinsic factors and exposures in 74%, 22%, 4% and 0% of

cases, respectively. Cellular composition explains > 25% of out-of-sample variance for 13% of CpG sites (n = 86,046; Figure 5A, C and Table S6), with the highest variance explained by cellular composition for one CpG site being 68.5%. The 16,034 CpG sites for which > 50% of variance is explained by cellular composition are typically located in genes related to the immune system (Top 3 gene ontology terms: leukocyte activation, cell activation, cell activation involved in immune response, $P_{adj} < 1.0\times10^{-28}$; Table S4). These CpG sites are concentrated in enhancer regions (95% CI: [3.41 3.69]; Figure S6B), and largely depleted from TSS (95% CI: [0.0761 0.106]), reflecting the importance of enhancer DNA methylation in cell-type identity.

For the 2,521 CpG sites where the model explains > 75% of variance, local genetic variation is the strongest predictor in 99% of cases (Figure 5C and Table S6). Local genetic variation explains > 25% of DNA methylation variance at 23,796 CpG sites, and almost as many when adjusting for cellular composition (n = 23,062) (Figure 5A, B), indicating that genetic effects on 5mC levels are cell-composition-independent. Intrinsic factors explain > 25% of out-of-sample variance at 3,621 CpG sites, and > 75% at 17 sites (Figure 5C). When conditioning on cell composition, these numbers dropped to 379 and 7 CpG sites, respectively, suggesting that the predictive ability of age and sex is partly mediated by immune cell composition (Figure 5B). Interestingly, environmental exposures are the weakest predictor of 5mC levels, explaining > 25% of the variance at only 23 CpG sites and with a maximum variance explained for a CpG site of 51%.

Finally, we estimated the proportion of variance explained by genotype × age, genotype × sex and genotype × exposure interactions, by considering the difference of the out-of-sample variance explained by models including interaction terms and models with only main effects (**STAR Methods**). We found a significant increase in predictive ability when including interaction terms for 1,984 CpG sites (ANOVA $P_{adj} < 0.05$). However, the effects were typically modest: only 35 CpG sites showed an increase in the proportion of variance explained larger than 4% (Figure 5B). The largest difference was found for a CpG site in the TSS of the *ENOSF1* gene, where the interaction model explained an additional 11.1% of DNA methylation variance (Table S6). Collectively, these results show that cellular composition and local genetic variation are the main drivers of DNA methylation variation in the blood of adults, reinforcing the critical need to study epigenetic risk factors and biomarkers of disease in the context of these factors.

DISCUSSION

Here, we present a rich data resource that delineates the contribution of genetics, age, sex, environmental factors, cellular composition and their interactions to variation in the DNA methylome. All the results can be explored via a web-based browser (MIMETH browser), to facilitate the exploration of the estimated effects of these factors on DNA methylation variation. We show that genetic variation controlling 5mC levels is likely to affect phenotype variation and disease risk, and often controls the expression of TFs. Furthermore, the remote genetic control of DNA methylation is driven by variants nearby *ZNF* genes, consistent with a role of ZFPs as direct regulators of 5mC levels (Marchal and Miotto, 2015). Furthermore, we show that remote meQTLs preferentially affect 5mC levels of *ZNF* genes, supporting the view that the major targets of ZFPs-mediated regulation are *ZNF* genes themselves (O'Geen et al., 2007). Most ZFPs possess a Krüppel-associated box (KRAB) domain, a DNA-binding domain that elicits KAP1-mediated transcriptional repression and induce heterochromatin by recruiting chromatin remodelers and DNA methyltransferases (Quenneville et al., 2012; Vogel et al., 2006; Zuo et al., 2012), providing a putative mechanism for the direct regulation of DNA methylation by KRAB-ZFPs. This is also supported by the widespread effect of a *SENP7* regulatory variant on 5mC levels of a *KRAB-ZNF* gene cluster on chromosome 19; SENP7 is a SUMO protease involved in the deSUMOylation of KAP1 that allows its chromatin remodelling activity (Garvin et al., 2013).

Our study reveals three different biological mechanisms underlying age-related changes in DNA methylation. The first elicits increased 5mC variance with age and is related to epigenetic drift (Fraga et al., 2005; Jones et al., 2015), likely caused by the progressive decline in fidelity of the DNA methylation maintenance machinery. The second produces cell-composition-independent, global DNA demethylation and CGI-associated hypermethylation. Age-associated DNA demethylation could be related to the downregulation of DNMT3A/B *de novo* methyltransferases, whereas CGI-associated hypermethylation may result from the downregulation of the Polycomb repressive complexes 1 and 2 and/or TET proteins, coupled with a loss of H3K27me3 marks (Beerman et al., 2013; Li et al., 2018; Williams et al., 2011). Alternatively, these changes may be related to the mitotic clock, which assumes a progressive accumulation of DNA methylation changes with mitotic divisions, including loss of methylation at partially methylated domains (PMD) and gain of methylation at PRC2-marked CpG-rich regions (Kim et al., 2005; Yang et al., 2016; Zhou et al., 2018). Both scenarios are supported by the enrichment of Polycomb-repressed regions in age-associated CpG sites, and of binding sites of PRC-related TFs in CpG sites methylated with age. The third mechanism elicits cell-composition-mediated demethylation at all compartments of the epigenome, particularly at enhancers of myeloid activation genes. This process likely reflects an

increased degree of differentiation in the lymphoid compartment. Single-cell methylomes of differentiating and dividing white blood cells will help determine the role of mitotic and post-mitotic 5mC changes during epigenetic aging.

Latent infections are known to profoundly alter the number, activation status and transcriptional profiles of immune cell populations, yet their epigenetic consequences have attracted little attention. We found that CMV infection elicits widespread changes in the blood DNA methylome, in contrast with other herpesviruses such as EBV, HSV-1, HSV-2 and VZV. We observe that most CMV effects are mediated by the profound changes in blood cell composition caused by CMV (Patin et al., 2018), including the inflation of CMV-specific memory $CD8^+$ T cells (Klenerman and Oxenius, 2016). However, we also detected cell-composition-independent effects of CMV infection, suggesting that the herpesvirus can directly regulate the host epigenome. Methylated CpG sites in CMV^+ donors are targeted by BRD4, a key host regulator of CMV gene expression and latency (Groves et al., 2021), suggesting that this TF, when upregulated during latent CMV infection, binds both viral and host genomes. Furthermore, CMV^+ donors are characterized by a strong increase in 5mC levels at *LTBP3*, the product of which is involved in TGF-β secretion. TGF-β is a well-known immunosuppressive cytokine induced by CMV infection (Mason et al., 2012), which represents a possible strategy of the virus to escape host immunity. These results suggest that the capacity of CMV to manipulate the host epigenetic machinery results in epigenetic changes of latently infected cells.

Another interesting finding of our study is that environmental exposures explain a small fraction of the variance of DNA methylation in healthy adults, at odds with the common view that the epigenome is strongly affected by the environment (Feil and Fraga, 2012). Twin studies have estimated the heritability of DNA methylation to range from ~20-40% (Bell et al., 2012; Grundberg et al., 2012; van Dongen et al., 2016), suggesting that environmental effects, along with gene × environment interactions, account for the remaining 60-80% (Teschendorff and Relton, 2018). However, other factors, including cellular composition and measurement error, may account for most of the unexplained variance. Consistently, we estimated that cellular composition explains >25% of the variance for ~13% of the DNA methylome, and it has been estimated that measurement error may explain >50% (Li et al., 2017). Nevertheless, a limitation of our study is that perinatal and early life exposures, which are thought to contribute extensively to epigenetic variation in adulthood (Feil and Fraga, 2012), have not been extensively assessed in the *Milieu Intérieur* cohort. In addition, it has been hypothesized that gene × environment interactions are central to understand the role of epigenetics in development (Boyce and Kobor, 2015), but statistical evidence for interaction effects requires larger cohorts (Fleiss, 2011), suggesting that our results might represent the small, perceptible fraction of a large number of weak effects (Czamara et al., 2019; Teh et al., 2014). Large, longi-

tudinal cohorts addressing the developmental origins of disease are needed to shed new light on the role of DNA methylation in the interplay between genes and the environment.

Collectively, our findings have broad consequences for the study and interpretation of epigenetic factors involved in disease risk. First, a third of the DNA methylome is affected by genetic variants, some of which are associated with disease risk. Epigenetic associations with a given disease or trait may thus result from the pleiotropic effects of genetic variants on DNA methylation, which may confound interpretation. Second, because age, sex, CMV infection, smoking and chronic low-grade inflammation influence disease risk (Furman et al., 2019; Mauvais-Jarvis et al., 2020; Niccoli and Partridge, 2012; Samet, 2013; Savva et al., 2013), our results highlight the critical need to consider such factors in EWAS. Third, our analyses clearly show that the effects of age, CMV serostatus and CRP levels are largely mediated by fine-grained changes in immune cell proportions. This reinforces the view that EWAS must be interpreted with caution, particularly when standard corrections using *estimated* cell proportions (Houseman et al., 2012; Koestler et al., 2016; Teschendorff et al., 2017) are incomplete. The integration of DNA methylation profiling and fine-grained measurements of immune cell subsets, such as the data used here, could also help improving the estimation of blood cell composition from DNA methylation and corrections for cellular heterogeneity. Finally, our findings highlight the major epigenetic impact of aging, persistent viral infections and inflammation through fine-grained changes in blood cell proportions, highlighting the need to assess the respective role of DNA methylation and altered cellular composition in the etiology of disease (Lappalainen and Greally, 2017). Large-scale studies using single-cell approaches will help overcome these challenges, and are anticipated to further decode the epigenetic mechanisms underlying healthy aging and the environmental causes of human disease.

ACKNOWLEDGMENTS

We thank Sarah Merrill, Nicole Gladish, Violaine Saint-André, Lucas Husquin and the *Milieu Intérieur* scientific advisory board for helpful discussions. We acknowledge the help of the HPC Core Facility of Institut Pasteur for this work. This work benefited from support of the French government's program 'Investissement d'Avenir', managed by the *Agence Nationale de la Recherche* (reference 10-LABX-69-01).

AUTHOR CONTRIBUTIONS

L.Q.-M. initiated the study. J.B., E.P. and L.Q.-M. conceived and developed the study. A.U. prepared DNA samples. D.T.S.L., J.L.M. and M.S.K. acquired Illumina Infinium MethylationEPIC array data. J.B. performed all analyses, with contributions from S.A.K.A., K.T. and E.P.. E.P. supervised all analyses. A.J. developed the MIMETH web browser. D.D. and M.L.A. advised on experiments. M.R., M.S.K., D.D. and M.L.A. advised on data interpretation. J.B., E.P. and L.Q.-M. wrote the manuscript. All authors discussed the results and contributed to the final manuscript.

The *Milieu Intérieur* Consortium[¶] is composed of the following team leaders: Laurent Abel (Hôpital Necker), Andres Alcover, Hugues Aschard, Philippe Bousso, Nollaig Bourke (Trinity College Dublin), Petter Brodin (Karolinska Institutet), Pierre Bruhns, Nadine Cerf-Bensussan (INSERM UMR 1163 – Institut Imagine), Ana Cumano, Christophe d'Enfert, Ludovic Deriano, Marie-Agnès Dillies, James Di Santo, Françoise Dromer, Gérard Eberl, Jost Enninga, Jacques Fellay (EPFL, Lausanne), Ivo Gomperts-Boneca, Milena Hasan, Gunilla Karlsson Hedestam (Karolinska Institutet), Serge Hercberg (Université Paris 13), Molly A. Ingersoll, Olivier Lantz (Institut Curie), Rose Anne Kenny (Trinity College Dublin), Mickaël Ménager (INSERM UMR 1163 – Institut Imagine) Hugo Mouquet, Cliona O'Farrelly (Trinity College Dublin), Etienne Patin, Sandra Pellegrini, Antonio Rausell (INSERM UMR 1163 – Institut Imagine), Frédéric Rieux-Laucat (INSERM UMR 1163 – Institut Imagine), Lars Rogge, Magnus Fontes, (Institut Roche), Anavaj Sakuntabhai, Olivier Schwartz, Benno Schwikowski, Spencer Shorte, Frédéric Tangy, Antoine Toubert (Hôpital Saint-Louis), Mathilde Touvier (Université Paris 13), Marie-Noëlle Ungeheuer, Christophe Zimmer, Matthew L. Albert (insitro)§, Darragh Duffy[§], Lluis Quintana-Murci[§]

[¶] Unless otherwise indicated, partners are located at Institut Pasteur, Paris; [§] Co-coordinators of the Milieu Intérieur Consortium; Additional information can be found at: https://milieuinterieur.fr/en/

DECLARATION OF INTERESTS

M.L.A. and A.U. are full-time employees of insitro. Other authors declare no competing interests.

FIGURES AND LEGENDS

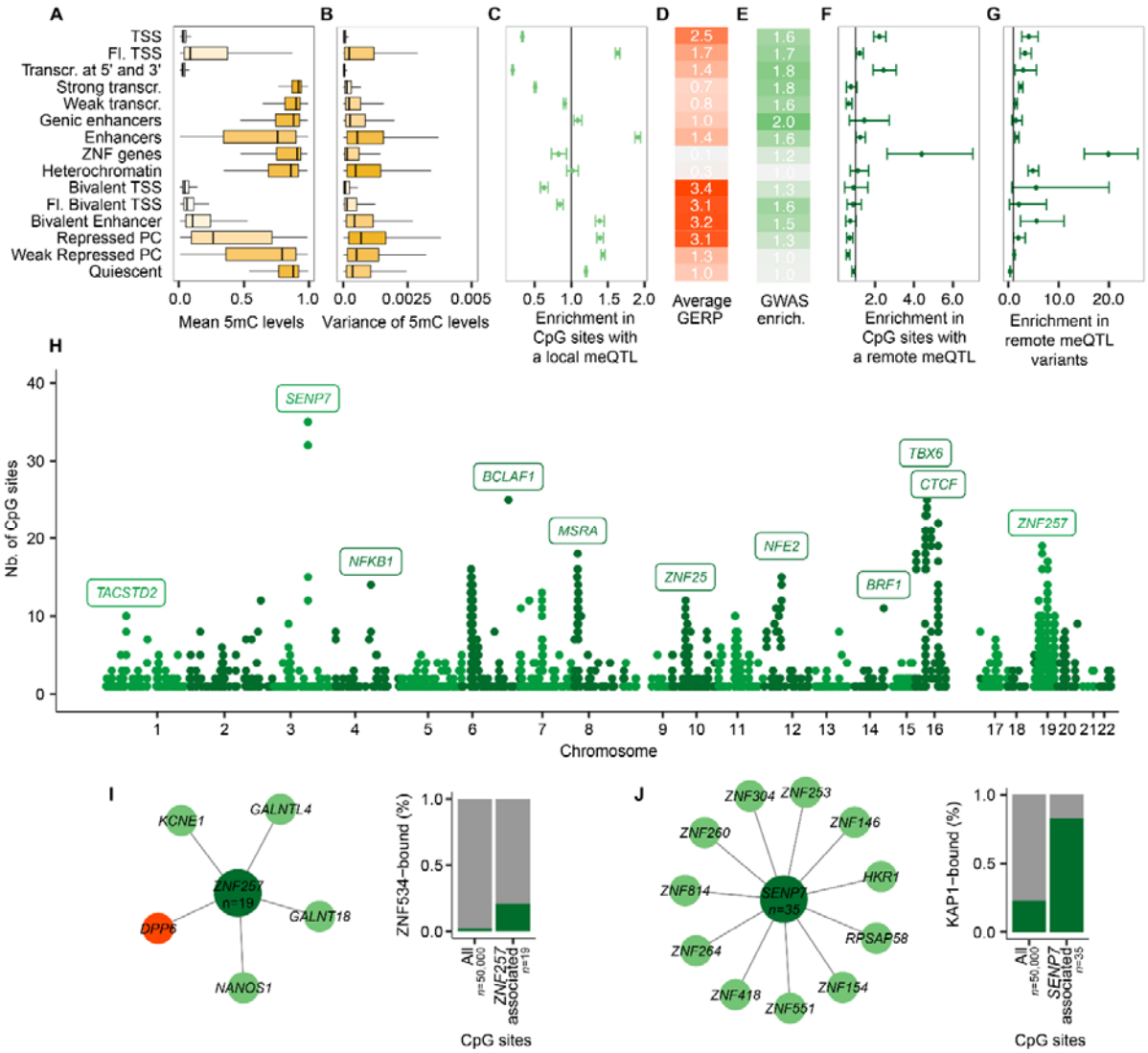

Figure 1. Local and Remote Genetic Control of DNA Methylation Variation in Whole Blood

(A) Distributions of 5mC levels at 644,517 CpG sites averaged over 958 donors, across 15 chromatin states.

(B) Distributions of the variance of 5mC levels at 644,517 CpG sites among 958 donors, across 15 chromatin states.

(C) Enrichment in CpG sites associated with local meQTL variants, across 15 chromatin states.

(D) Average Genomic Evolutionary Rate Profiling (GERP) scores, across 15 chromatin states.

(E) Enrichment of local meQTL variants in disease/trait-associated variants, across 15 chromatin states.

(F) Enrichment in CpG sites associated with remote meQTL variants, across 15 chromatin states.

(G) Enrichment in remote meQTL variants, across 15 chromatin states.

(H) Number of associated CpG sites per remote variant. Variants are annotated based on their closest gene and only one variant per gene is shown.

(I) Network of genes at which CpG sites are associated remotely with *ZNF257* genetic variation. Green and orange colors denote positive and negative effects. Bar plots denote the proportion of CpG sites that overlap with binding sites of the ZNF534 TF.

(J) Network of genes at which CpG sites are associated remotely with *SENP7* genetic variation. Bar plots denote the proportion of CpG sites that overlap with binding sites of the KAP1 TF.

(A-G) Chromatin states were defined in $CD4^+$ naïve T cells (Roadmap Epigenomics et al., 2015). TSS, Fl. and PC denote transcription start site, flanking and Polycomb, respectively.

(C, F, G) The odds-ratio and 95% CI are indicated by the point and error bars, respectively.

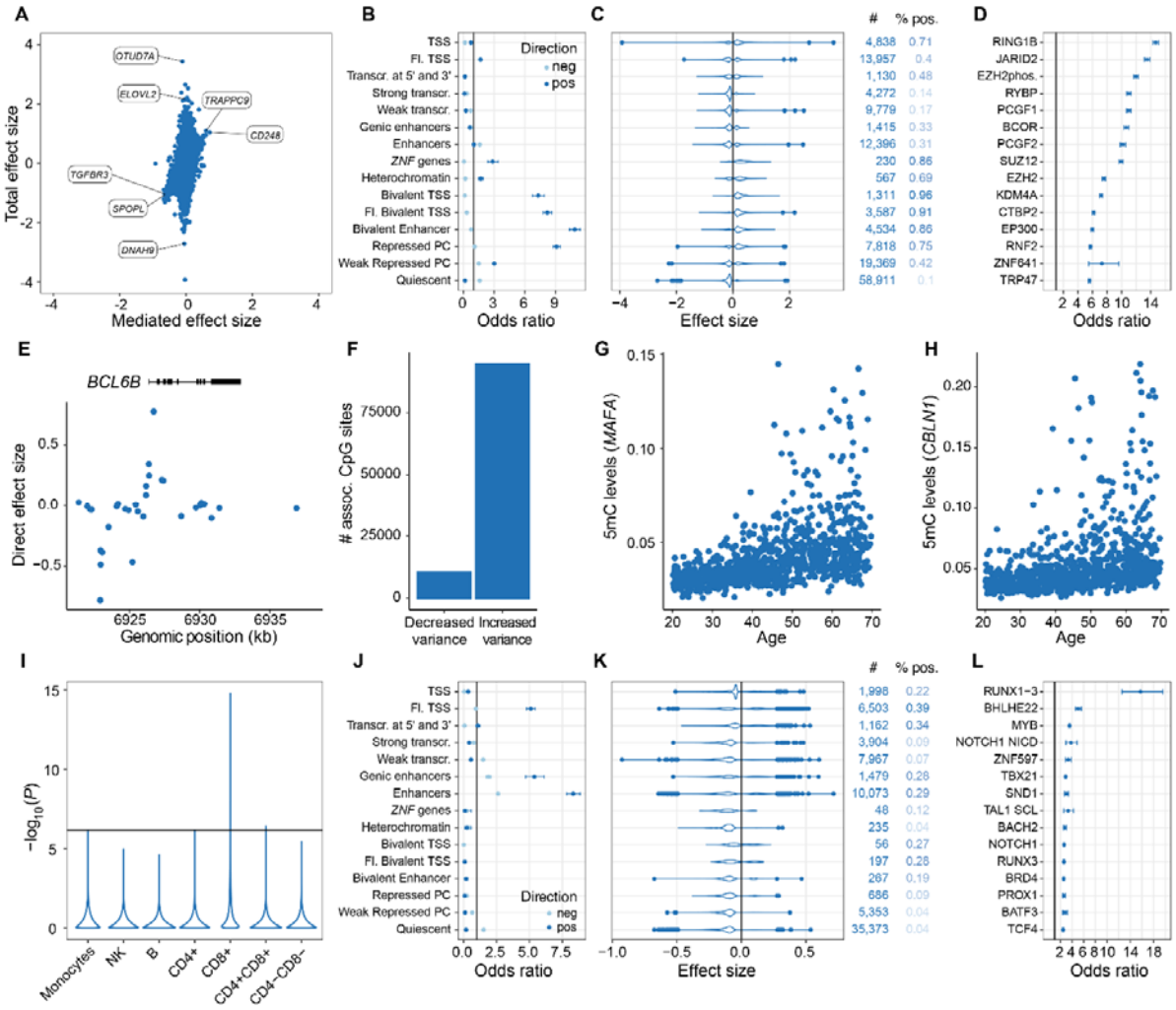

Figure 2. Direct and Cell-Composition-Mediated Effects of Aging on the Blood DNA Methylome

(A) Total effects against cell-composition-mediated effects of age on 5mC levels (50 year effect). Only CpG sites with a significant total and/or cell-composition-mediated effect are shown. Labels denote genes with strong total or cell-composition-mediated effects of age.

(B) Enrichment in CpG sites with significant direct effects of age, across 15 chromatin states.

(C) Distributions of significant direct effects of age, across 15 chromatin states. Numbers on the right indicate the number of associated CpG sites and proportion of positive effects.

(D) Enrichment of CpG sites with a significant positive, direct effect of age in binding sites for TFs. The 15 most enriched TFs are shown, out of 1,165 tested TFs.

(E) Genomic distribution of direct age effects at the *BCL6B* locus.

(F) Number of CpG sites with a significant decreased or increased variance with age.

(G) Increased variance of 5mC levels with age at the *MAFA* locus.

(H) Increased variance of 5mC levels with age at the *CBLN1* locus.

(I) *P*-value distributions for the effect of age on 5mC levels within six major immune cell types, compared to the effect of age within neutrophils.

(J) Enrichment of CpG sites with significant cell-composition-mediated effects of age, across 15 chromatin states.

(K) Distributions of significant cell-composition-mediated effects of age, across 15 chromatin states. Numbers on the right indicate the number of associated CpG sites and proportion of positive effects.

(L) Enrichment of CpG sites with significant cell-composition-mediated, positive effects of age in binding sites for TFs. The 15 most enriched TFs are shown, out of 1,165 tested TFs.

(B, D, J, L) The odds-ratio and 95% CI are indicated by the point and error bars, respectively.

(A, C, E, K) Effect sizes are given in the M value scale.

(G, H) 5mC levels are given in the β value scale.

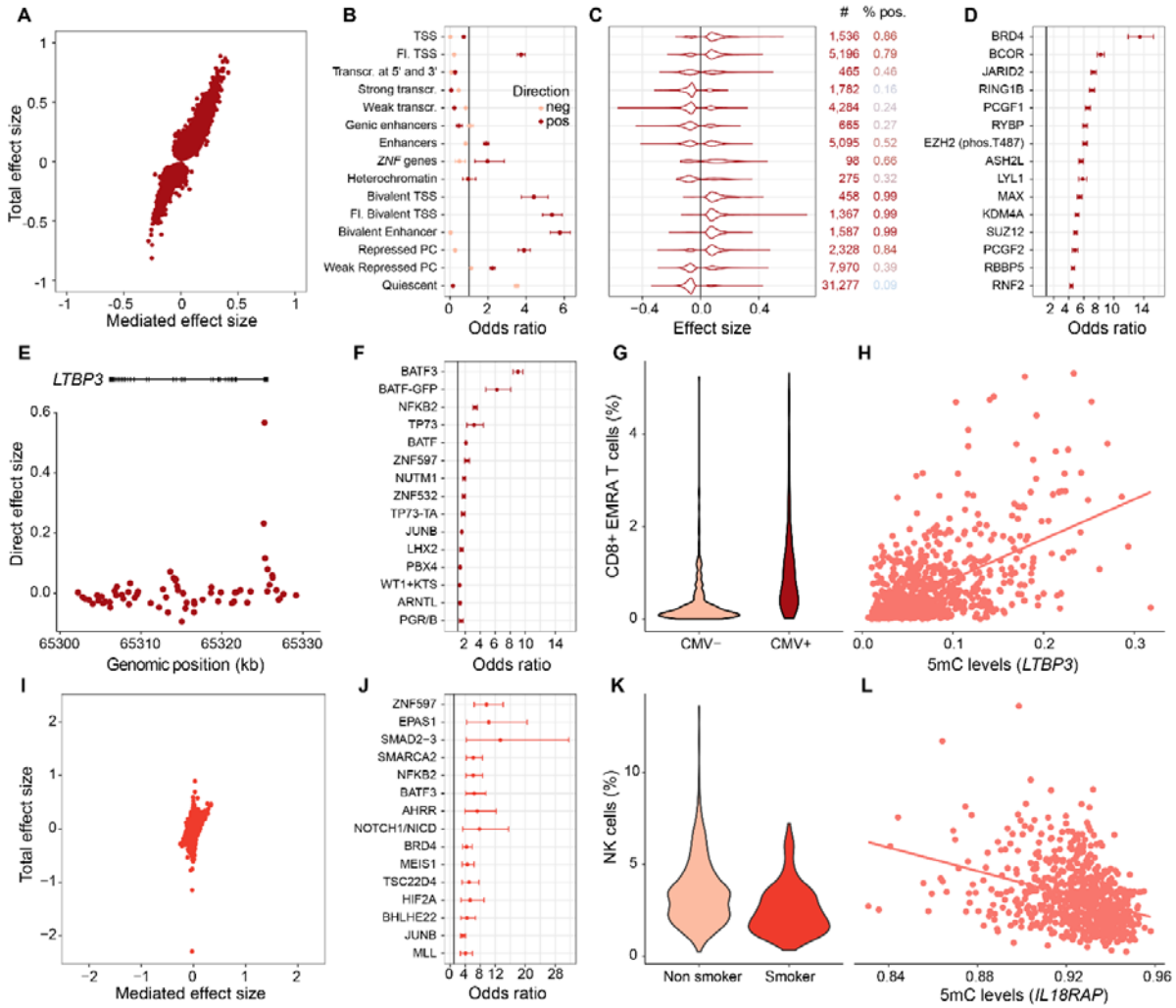

Figure 3. Effects of Latent Cytomegalovirus Infection and Smoking on the Blood DNA Methylome

(A) Total effects against cell-composition-mediated effects of latent CMV infection on 5mC levels.

(B) Enrichment in CpG sites with a significant direct effect of CMV infection, across 15 chromatin states.

(C) Distributions of significant direct effects of CMV infection across 15 chromatin states. Numbers on the right indicate the number of associated CpG sites and proportion of positive effects.

(D) Enrichment of CpG sites with a significant direct, positive effect of CMV infection in binding sites for TFs. The 15 most enriched TFs are shown, out of 1,165 tested TFs.

(E) Genomic distribution of direct effects of CMV infection at the *LTBP3* locus.

(F) Enrichment of CpG sites with a significant direct, negative effect of CMV infection in binding sites for TFs. The 15 most enriched TFs are shown, out of 1,165 tested TFs.

(G) Distributions of the proportion of $CD8^+$ T_{EMRA} cells in CMV^- and CMV^+ donors.

(H) 5mC levels at the *LTBP3* locus against the proportion of $CD8^+$ T_{EMRA} cells.

(I) Total effects against cell-composition-mediated effects of smoking status on 5mC levels.

(J) Enrichment of CpG sites with significant direct, positive effects of smoking in binding sites for TFs. The 15 most enriched TFs are shown, out of 1,165 tested TFs.

(K) Distributions of the proportion of NK cells in non-smokers and smokers.

(L) 5mC levels at the *IL18RAP* locus against the proportion of NK cells.

(B, D, F, J) The odds-ratio and 95% CI are indicated by the point and error bars, respectively.

(A, C, E, I) Effect sizes are given in the M value scale.

(H, L) 5mC levels are given in the β value scale.

(A, I) Only CpG sites with a significant total and/or cell-composition-mediated effect are shown.

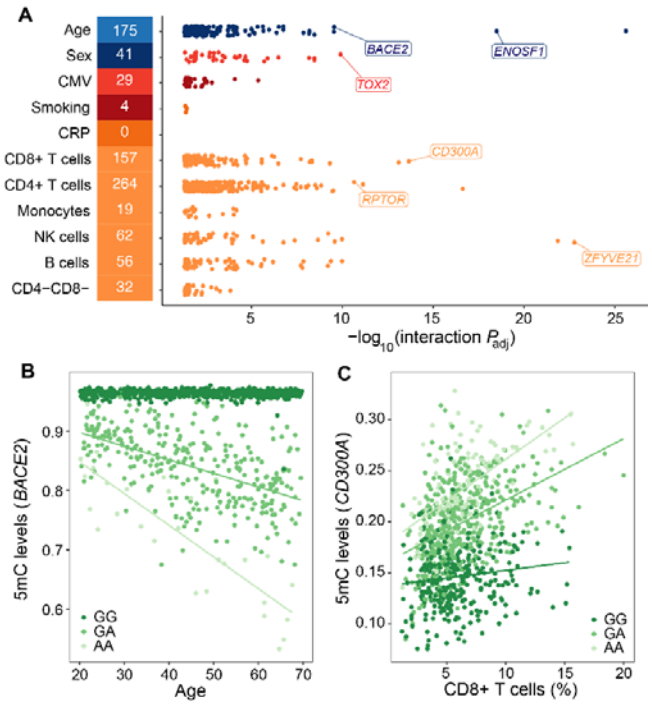

Figure 4. Effects of Gene × Environment Interactions on the Blood DNA Methylome

(A) *P*-value distributions for significant effects of genotype × age, genotype × sex, genotype × exposures and genotype × cell type interactions. The number of significant associations is indicated on the left. Labels denote genes with strong interaction effects.

(B) Genotype-dependent effect (rs2837990 variant) of age on 5mC levels at the *BACE2* locus.

(C) $CD8^+$ T cell-dependent effect of the rs12939435 variant on 5mC levels at the *CD300A* locus.

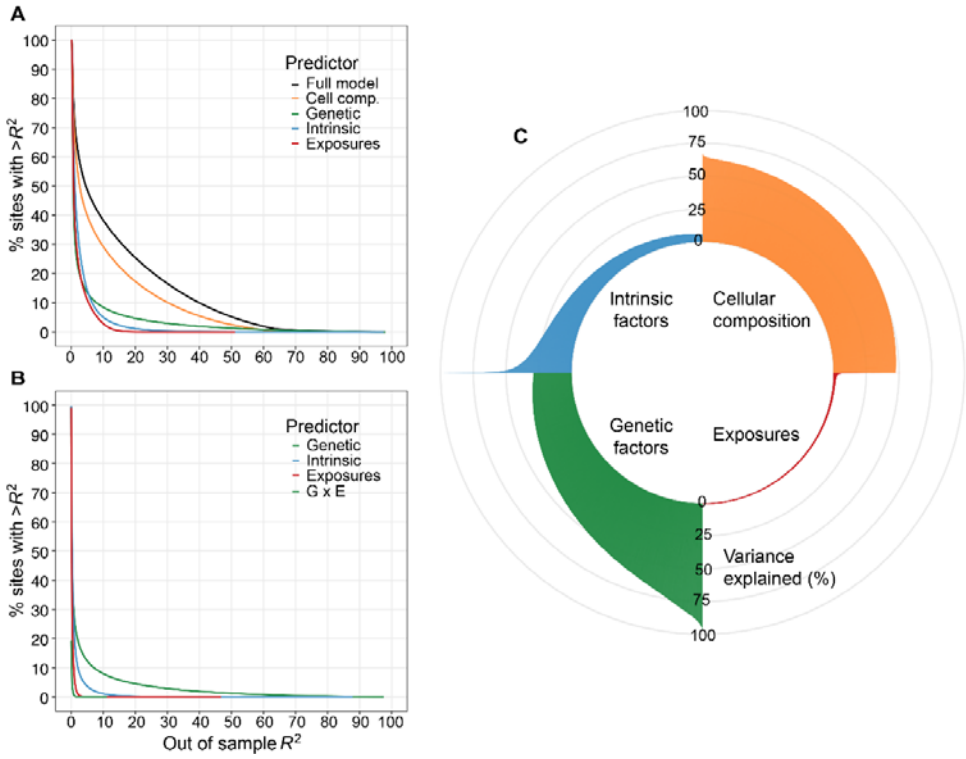

Figure 5. Best Predictors of the Blood DNA Methylome of Adults

(A) Complementary cumulative distribution function of the out-of-sample variance explained by the full model, blood cell composition, genetic factors, intrinsic factors and exposures, for 644,517 CpG sites.

(B) Complementary cumulative distribution function of the out-of-sample variance explained by genetic factors, intrinsic factors, exposures and gene × environment (G × E) interactions, when conditioning on blood cell composition, for 644,517 CpG sites.

(C) Proportion of the explained out-of-sample variance of 5mC levels for the 20,000 CpG sites with the variance most explained by blood cell composition, genetic factors, intrinsic factors and exposures, respectively.

TABLES

DNA Sequence variant	Chr.	Position	Closest gene	eQTL *P*-value	eQTL direction	#CpG sites	Positive effects (%)	TF	TFBS enrichment [95% CI]
rs77081633	6	136589425	*BCLAF1*	-	-	25	48%	BCLAF1	7.4 [3.04, 18]
rs11850055	14	105754532	*BRF1*	1.1×10^{-22}	Positive	11	0%	BRF1	1011.3 [162.91, 4300]
rs60626639	16	67625797	*CTCF*	-	-	22	96%	CTCF	16.8 [2.71, 694]
rs11986122	8	10009949	*MSRA*	1.3×10^{-177}	Negative	16	56%	-	-
rs79755767	12	54698408	*NFE2*	-	-	14	100%	NFE2	141 [21.16, 5765]
rs1585215	4	103444474	*NFKB1*	8.9×10^{-45}	Positive	14	0%	NFKB1	59.8 [15.79, 337]
rs12491955	3	101146597	*SENP7*	4.2×10^{-302}	Positive	35	9%	KAP1	16.3 [6.66, 48]
rs10889104	1	59046496	*TACSTD2*	4.2×10^{-302}	Positive	10	0%	-	-
rs3809627	16	30103160	*TBX6*	3.5×10^{-135}	Negative	26	31%	-	-
rs1005278	10	38218748	*ZNF25*	8.9×10^{-171}	Positive	10	100%	-	-
rs10417143	19	22373303	*ZNF257*	2.8×10^{-70}	Positive	19	16%	ZNF534	5.5 [0.62, 23]

Table 1. Genetic variants with multiple, remote effects on the blood DNA methylome. DNA sequence variants that affect 5mC levels at more than 10 remote CpG sites can control mRNA levels of a nearby TF or chromatin remodeler, and remotely associated CpG sites can be enriched in binding sites for the corresponding TF, or a TF related to it. Chr. denotes the chromosome where the DNA sequence variant is located. eQTL *P*-value is the *P*-value of association between the master variant and mRNA levels of a close gene (Võsa et al., 2018). #CpG sites denote the number of CpG sites remotely associated with the DNA sequence variant. TFBS stands for TF binding sites.

Factor	Associated CpG sites (total effect)	Associated CpG sites (direct effect)	Associated CpG sites (mediated effect)
age	258,830	144,114	75,301
sex	186,545	126,904	44,667
CMV serostatus	233,014	64,383	217,223
smoking status	7,257	2,416	20,381
log CRP levels	20,043	480	-
log triglyceride levels	4	9	-
years since last cigarette	10	8	-
years of cigarette smoking	0	7	-
surgery	6	6	-
abdominal circumference	1	6	-
log HDL levels	1	3	-
metabolic score	1	3	-
raw fruit consumption	2	2	-
log protein levels	1	2	-
weight	0	2	-
heart rate	76,018	0	-
auricular temperature	59,728	0	-
log CMV IgG levels	52,564	0	-
hour of sampling	38,884	0	-
log glycaemia levels	3	0	-
log chloride levels	2	0	-

Table 2. Number of CpG sites significantly associated with intrinsic factors and exposures. Out of 141 tested factors, 20 and 16 have a significant total effect (i.e., cell-composition-independent and cell-composition-mediated effects) or a direct effect (i.e., cell-composition-independent effect) on the 5mC levels of more than one CpG site, respectively. “log” denotes log-transformation, CMV: cytomegalovirus, CRP: C-reactive protein, and HDL: high-density lipoprotein.

STAR METHODS

- KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Whole blood DNA from 979 healthy donors	The *Milieu Intérieur* cohort, France	NCT03905993
Critical Commercial Assays		
Nucleon BACC3 DNA extraction kit	Ge Healthcare	RPN8512
Infinium MethylationEPIC array	Illumina	WG-317-1003
Deposited Data		
Milieu Intérieur MethylationEPIC data	European Genome-Phenome Archive	EGAS0000XXXXXXX
Milieu Intérieur genotype data	European Genome-Phenome Archive	EGAS00001002460
Milieu Intérieur flow cytometry data	European Genome-Phenome Archive	EGAS0000XXXXXXX
Software and Algorithms		
R-3.6.0	R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria	https://www.R-project.org/
minfi R package	PMID:28035024	https://doi.org/doi:10.18129/B9.bioc.minfi
irlba R package		https://cran.r-project.org/web/packages/irlba/index.html
missForest R package	PMID: 22039212	https://cran.r-project.org/web/packages/missForest/index.html
sva R package, ComBat function	PMID:16632515	https://rdrr.io/bioc/sva/man/ComBat.html
MatrixEQTL R package	PMID:22492648	https://cran.r-project.org/web/packages/MatrixEQTL/index.html
FlowSorted.Blood.EPIC R package	PMID:29843789	https://doi.org/doi:10.18129/B9.bioc.FlowSorted.Blood.EPIC
sandwich R package	(Stekhoven and Buhlmann, 2012)	https://cran.r-project.org/web/packages/sandwich/index.html

pbkrtest R package	(Halekoh and Højsgaard, 2014)	https://cran.r-project.org/web/packages/pbkrtest/index.html
missMethyl R package	PMID: 26424855	https://doi.org/doi:10.18129/B9.bioc.missMethyl

● CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and information should be directed to and will be fulfilled by the Lead Contact, Pr. Lluís Quintana-Murci (quintana@pasteur.fr).

● EXPERIMENTAL MODEL AND SUBJECT DETAILS

The *Milieu Intérieur* cohort

The *Milieu Intérieur* cohort was established with the goal to identify genetic variation and environmental exposures that affect phenotypes related to the immune system in the adult, healthy population. The 1,000 healthy donors of the *Milieu Intérieur* cohort were recruited by BioTrial (Rennes, France), and included 500 women and 500 men. Donors included 100 women and 100 men from each decade of life, between 20 and 69 years of age. Donors were selected based on various inclusion and exclusion criteria that are detailed elsewhere (Thomas et al., 2015). Briefly, donors were required to have no history or evidence of severe/chronic/recurrent pathological conditions, neurological or psychiatric disorders, alcohol abuse, recent use of illicit drugs, recent vaccine administration, and recent use of immune modulatory agents. To avoid the influence of hormonal fluctuations in women, pregnant and peri-menopausal women were not included. To avoid genetic stratification in the study population, the recruitment of donors was restricted to individuals whose parents and grandparents were born in Metropolitan France.

Ethical approvals

The study is sponsored by the Institut Pasteur (Pasteur ID-RCB Number: 2012-A00238-35) and was conducted as a single center study without any investigational product. The *Milieu Intérieur* clinical study was approved by the *Comité de Protection des Personnes* — Ouest 6 (Committee for the protection of persons) on June 13, 2012 and by the French *Agence Nationale de Sécurité du Médicament* (ANSM) on June 22, 2012. The samples and data used in this study were formally established as the *Milieu Intérieur* biocollection (study# NCT03905993), with approvals by the Comité de Protection des Personnes – Sud Méditerranée and the Commission nationale de l'informatique et des libertés (CNIL) on April 11, 2018.

● METHOD DETAILS

DNA sampling and extraction

Whole blood samples were collected from the 1,000 *Milieu Intérieur* healthy, fasting donors on Li-heparin, every working day from 8AM to 11AM, from September 2012 to August 2013, in Rennes, France. Tracking procedures were established in order to ensure delivery to Institut Pasteur (Paris) within 6 hours of blood draw, at a temperature between 18°C and 25°C. Upon receipt, samples were kept at room temperature until DNA extraction. DNA was extracted using the Nucleon BACC3 genomic DNA extraction kit (GE Healthcare, Illinois, USA). High-quality genomic DNA was obtained for 978 out of the 1,000 donors.

DNA methylation profiling and data quality controls

Extracted genomic DNA was treated with sodium bisulfite (Zymo Research, California, USA). Bisulfite-converted DNA was applied to the Infinium MethylationEPIC BeadChip (Illumina, California, USA), using the manufacturer's standard conditions. The MethylationEPIC BeadChip measures 5mC levels at 866,836 CpG sites in the human genome. Raw IDAT files were processed with the minfi R package (Fortin et al., 2017). All samples showed average detection *P*-values lower than 0.005. No sample showed a mean of methylated intensity signals lower than 3 standard deviations from the cohort average. Thus, no samples were excluded based on detection *P*-values or methylated intensity signals. The sex predicted from 5mC signals on sex chromosomes matched the declared sex for all samples (Figure S1A). Using the 59 control SNPs included in the MethylationEPIC array, a single sample showed high genotype discordance with the genome-wide SNP array data (see 'Genome-wide DNA genotyping' section) and was thus excluded (Figure S1B). Unmethylated and methylated intensity signals were converted to M-values. A total of 2,930 probes with >1% missingness (i.e., detection *P*-value > 0.05 for more than 1% of donors) were excluded and remaining missing data (missingness = 0.0038%) were imputed by mean substitution. Using the irlba R package, Principal Component Analysis (PCA) of M values identified nine outlier samples, including eight that were processed on the same array (Figure S1C), which were also excluded. The "noob" background subtraction method (Triche et al., 2013) was applied on M values for the remaining 969 samples, which showed highly consistent epigenome-wide 5mC profiles (Figure S1D,E).

To identify batch effects on the DNA methylation data, we searched for the factors that were the most associated with the top 20 PCs of the PCA of noob-corrected M values. We used a linear mixed model that included the proportion of lymphocytes, age, sex and cytomegalovirus (CMV) serostatus as fixed effects, and slide position and sample plate as random effects. Strong

associations were observed between the first four PCs and slide position and sample plate (Figure S1F, G). M values were thus corrected for these two batch effects using ComBat (Johnson et al., 2007). After ComBat correction, the ten first PCs of a PCA of M values were associated with factors known to affect DNA methylation, including blood cell composition, age and sex (Figure S1H-J), indicating no other, strong batch effect on the data. M-values were converted to β values, considering that $\beta = 2^M / (2^M + 1)$. Because outlier 5mC values due to measurement error could inflate the type I error rate of regression models, we excluded, for each CpG site, M or β values that were greater than 5 × standard deviations (SD) from the population average, corresponding to <0.1% of all measures. We also excluded (i) 83,380 non-specific probes that share >90% sequence identity with several genomic regions (see details in (Price et al., 2013)), (ii) 118,575 probes that overlap a SNP with MAF>1% in the *Milieu Intérieur* cohort or in European populations from the 1,000 Genomes project (Auton et al., 2015), (iii) 558 probes that were absent from the Illumina annotations version 1.0 B4 and (iv) 16,876 probes located on sex chromosomes. As a result, the final, quality-controlled data was composed of 968 donors profiled at 644,517 CpG sites.

Flow cytometry

Protocols, panels, staining antibodies and quality control filters used for flow cytometry analyses are detailed elsewhere (Patin et al., 2018). Briefly, immune cell proportions were measured using ten eight-color flow-cytometry panels. The acquisition of cells was performed using two MACSQuant analyzers, which were calibrated using MacsQuant calibration beads (Miltenyi, Germany). Flow cytometry data were generated using MACSQuantify software version 2.4.1229.1. The mqd files were converted to FCS compatible format and analyzed by FlowJo software version 9.5.3. A total of 110 cell proportions were exported from FlowJo. Abnormal lysis or staining were systematically flagged by trained experimenters. We removed outliers by using a scheme detailed previously (Patin et al., 2018). Briefly, we used a distance-based approach that, for each cell-type, removes observations in the right tail if the distance to the closest observation in the direction of the mean is larger than 20% of the range of the observations. Similarly, observations in the left tail were removed if the distance to the closest observation in the direction of the mean is more than 15% than the range the observations. We removed 22 observations in total, including a maximum of 8 observations for a single cell type (i.e., for the proportion of neutrophils). Finally, missing data were imputed using the random forest-based missForest R package (Stekhoven and Buhlmann, 2012).

Genome-wide DNA genotyping

Protocols and quality control filters for genome-wide SNP genotyping are detailed elsewhere (Patin et al., 2018). Briefly, all the 1,000 *Milieu Intérieur* donors were genotyped on both the HumanOmniExpress-24 and the HumanExome-12 BeadChips (Illumina, California, USA), which include 719,665 SNPs and 245,766 exonic SNPs, respectively. Average concordance rate between the two genotyping arrays was 99.9925%. The final data set included 732,341 high-quality polymorphic SNPs. After genotype imputation and quality-control filters, a total of 11,395,554 SNPs was further filtered for minor allele frequencies > 5%, yielding a data set composed of 1,000 donors and 5,699,237 SNPs for meQTL mapping. Ten pairs of first to third-degree related donors were detected with KING 1.9 (Manichaikul et al., 2010). Out of the 968 donors whose blood methylome was profiled, 958 unrelated donors were kept for subsequent analyses.

● QUANTIFICATION AND STATISTICAL ANALYSIS

Circulating immune cells

One of the key questions in this study is whether differences in 5mC levels observed with respect to different factors are due to epigenetic changes occurring within cells or if they in fact reflect changes in cell composition. To answer this question, we adjusted models on measured proportions of 16 major subsets of blood: naïve, central memory (CM), effector memory (EM) and terminally differentiated effector memory (EMRA) subsets of $CD4^+$ and $CD8^+$ T cells, $CD4^-CD8^-$ T cells, B cells, dendritic cells, natural killer (NK) cells, monocytes, neutrophils, basophils and eosinophils (Patin et al., 2018). We also investigated whether some factors affect 5mC levels differently within cellular subsets. To answer this question, we derived interaction models, where measured cell proportions interacted with the factor. Since these models showed inflated variance for small subsets, we used a reduced set of 7 immune cell types for this analysis: $CD4^+$ and $CD8^+$ T cells, $CD4^-CD8^-$ T cells, B cells, NK cells, monocytes and neutrophils.

Local meQTL mapping

Local meQTL mapping was performed using the MatrixEQTL R package (Shabalin, 2012). Association was tested for each CpG site and each SNP in a 100-Kb window around the CpG site, by fitting a linear regression model assuming an additive allele effect. Models included the set of 16 immune cell proportions (see above) as predictors. They also included factors we have previously identified to have a large impact on blood and its molecular characteristics: a nonlinear age term encoded by 3 degrees-of-freedom (DoF) natural splines, sex, smoker status, ex-smoker status and CMV serostatus (Patin et al., 2018). We also adjusted for the top two PCs of a PCA of the genotype data. We did not include more PCs because of the low population substructure observed in the

cohort (Patin et al., 2018). For the i:th individual and the p:th CpG site, let y_i^p be the measured 5mC levels on the M value scale, SNP_i^m the number of minor alleles of the m:th tested SNP for the CpG site and $f_{\beta_{Age}}(Age_i)$ a nonlinear age term of natural splines with corresponding parameter vector β_{Age}. Moreover, let the vector c_i be measurements of the 16 immune cell types for the i:th individual and β_c be the corresponding parameter vector. The additive allele effect of the SNP was estimated by the parameter β_m in the model,

$$y_i^p = \mu + SNP_i^m \beta_m + c_i^T \beta_c + f_{\beta_{Age}}(Age_i) + Woman_i \beta_{Woman} + Exsmoker_i \beta_{Exsmoker} + Smoker_i \beta_{Smoker} + CMV_i \beta_{CMV} + PC1_i \beta_{PC1} + PC2_i \beta_{PC2} + \varepsilon_i, \quad \textit{Eq. 1}$$

Where ε_i is a symmetrical zero-mean distribution with constant variance.

Long-range meQTL mapping

Testing all possible associations between 644,517 CpG sites and 5,699,237 SNPs would require performing 3,769 billion statistical tests. To reduce the number of tests, long-range meQTL mapping was conducted on a selection of 50,000 CpG sites with the highest residual variance in the model described in *Eq. 1*, but with *m* indexing in this case the most associated local SNP for each site. For each of the 50,000 selected CpG sites, we then fitted one model per SNP located outside of a 1-Mb window around the CpG site. For each SNP-CpG pair, we estimated the additive allele effect of the remote SNP using the model specified in *Eq. 1*.

Local and long-range meQTL mapping were adjusted for multiple testing by employing a two-stage hierarchical procedure designed for the structure of tested hypotheses (Peterson et al., 2016). Consider all performed hypothesis tests, $H_{p,m_p}, p = 1, \dots, N,\ m_p = 1, \dots, M_p$, where N is the number of CpG sites and M_p is the number of SNPs considered for the p:th CpG site, with corresponding P-values, P_{p,m_p}. Define the family of all hypothesis tests performed for the p:th CpG site,

$$H_p = \left\{H_{p,1}, \dots, H_{p,M_p}\right\}.$$

For each such family, we tested the intersection hypothesis of no genetic control of 5mC levels at the CpG site. The *P*-value for this hypothesis was computed as the smallest Bonferroni-adjusted *P*-value in the set of *P*-values for the family,

$$P_p = \min_{m_p} \left\{ M_p P_{p,m_p} \right\}.$$

To adjust for multiplicity due to the number of CpG sites, we adjusted the *P*-value collection $P_p, p = 1, .., N$ by the Benjamini-Hochberg procedure. We considered an intersection hypothesis to be rejected if its Benjamini-Hochberg adjusted *P*-value was below 0.05. Let S be the number of rejected intersection hypotheses. In the final stage, we performed hypothesis tests for association of SNPs with 5mC levels of CpG sites under genetic control, *i.e.*, within families with a rejected intersection hypothesis. We considered a hypothesis H_{p,m_p} within a selected family to be rejected if,

$$M_p P_{p,m_p} \frac{N}{S} < 0.05.$$

This procedure controls the false discovery rate (FDR) for discovery of CpG sites under genetic control, the global FDR over all tests and the average family-wise error rate (FWER) over selected families (Peterson et al., 2016).

Detection of independent remote meQTLs

We designed the following scheme to compute a set Φ of independently associated remote SNPs for each CpG site, where all such SNPs are associated with 5mC levels y_p at the p:th CpG site conditional on the most associated local SNP and other SNPs in Φ. Define X_1 to be the set of SNPs with a long-range association to y_p and let x_0 be the most associated significant local SNP, if it exists. The set X_1 includes many SNPs that are in linkage disequilibrium (LD). The algorithm uses an iterative procedure to build sets M_j of SNPs, where in the j:th iteration, SNPs that are not associated with 5mC levels at the CpG site conditional on SNPs included in M_{j-1} are discarded, while the most associated is retained in M_{j}. In the final step, the set Φ is constructed by elements of M_{j} that are associated with 5mC levels at the CpG site conditional on all the other elements. Intuitively, Φ consists of the most associated SNP in each LD block. The algorithm is given in pseudocode in *Algorithm 1*, where the condition $\beta \neq 0$ is determined by an F test on the level $\alpha = 10^{-6}$.

Algorithm 1: Forming a set of long-range independently associated SNPs with a CpG site

If the CpG site is under local genetic control then let $M_1 = x_0$, otherwise let $M_1 = \emptyset$

Repeat for $j = 1, 2, ...$

$P = \{x \in X_j \setminus M_j\colon \beta_x \neq 0 \text{ in } y_p = \mu + x\beta_x + \sum_{z \in M_j} z\beta_z + \varepsilon, \varepsilon \sim (0, \sigma^2)\}$

If $P = \emptyset$ Exit

$X_{j+1} = P$

$\mathrm{M}_{j+1} = \mathrm{M}_j \cup \{x\colon x \text{ SNP with the smallest } P\text{-value in } P\}$

End

$\Phi = \{x \in M_{j+1} \setminus x_0\colon \beta_x \neq 0 \text{ in } y_p = \mu + x\beta_x + \sum_{z \in M_{j+1} \setminus \{x\}} z\,\beta_z + \varepsilon, \varepsilon \sim (0, \sigma^2)$

Epigenome-wide association studies of non-genetic factors

We assessed the effect of 141 non-genetic variables (Table S1) on the blood methylome of adults. The measured 5mC levels at a CpG site is an average of the methylation state of this CpG site in all cells in the blood sample. Cell composition is unlikely to have a strong causal effect on most of the investigated variables, with few exceptions, such as C-reactive protein (CRP) levels. However, many of the 141 candidate variables are likely to influence cell composition, which will cause a corresponding change in 5mC levels. We denote this effect the "(cell-type-)mediated effect". In addition, the variable might alter 5mC levels within individual cells, or within cell types. We denote this effect the "direct effect" (See Figure S3H for a schematic directed acyclic graph of the system). Several important factors are known to have a large effect on blood cell composition in healthy donors, the most important being age, sex, CMV serostatus and smoking (Patin et al., 2018). As an added complexity, these factors are also associated with most of the other variables in the study. Based on this framework, we investigated four questions, each one targeted by a separate statistical model.

The total effect

The total effect includes both changes in 5mC levels induced by changes in cellular composition and those induced within cell types. For each variable of interest x and CpG site pair, the total effect was estimated in a regression model including 5mC levels of the CpG site on the M value scale as response variable and x, a nonlinear age term of 3 DoF natural splines, sex, CMV serostatus, smoking status, the most associated significant local SNP, independently associated remote SNPs and the two first PCs of the genotype matrix as predictors. In addition, since we noticed variability

in 5mC levels across days of blood draw, we included date of blood draw as a random effect. Let j be the day of blood draw for the i:th individual. For the p:th CpG site, let y_i^p be the 5mC levels of the i:th individual on the M value scale, $f_{\beta_{Age}}(Age_i)$ a nonlinear age term of 3 DoF natural splines and SNP_i a vector of the number of minor alleles of independently associated SNPs with corresponding parameter vector β_{SNP}. The total effect of the variable x was estimated by the corresponding parameter β_x in the model,

$$\begin{aligned} y_i^p = \mu + x_i\beta_x + f_{\beta_{Age}}(\mathrm{Age_i}) + \mathrm{Woman}_i\beta_{\mathrm{Woman}} + \mathrm{Exsmoker}_i\beta_{\mathrm{Exsmoker}} \\ + \mathrm{Smoker}_i\beta_{\mathrm{Smoker}} + \mathrm{CMV}_i\beta_{\mathrm{CMV}} + \mathrm{PC1}_i\beta_{\mathrm{PC1}} + \mathrm{PC2}_i\beta_{\mathrm{PC2}} \\ + \mathrm{SNP}_i^T\beta_{\mathrm{SNP}} + \mathrm{DayOfSampling}_{j(i)} + \varepsilon_i, \end{aligned} \qquad \textit{Eq. 2}$$

where $\mathrm{DayOfSampling}_{j(i)} \sim \mathcal{N}(0, \sigma_d^2)$ and $\varepsilon_i \sim (0, \sigma^2)$. Aging was tested by removing x and replacing the non-linear age term with a linear one in *Eq. 2*. The effects of sex, smoking status and CMV serostatus were tested by removing x in *Eq. 2*. For variables concerning women only (e.g., age of menarche), we excluded men from the analysis and removed the $Woman_i\beta_{Woman}$ term. Hypothesis tests were performed by the Kenward-Roger approximation of the F-test for linear mixed models, implemented in the pbkrtest R package (Halekoh and Højsgaard, 2014).

The direct effect

Let the vector c_i be measurements of the 16 immune cell types for the i:th individual and β_c be the corresponding parameter vector. Using the same notation as for the total effect, the direct effect of the variable x was estimated by β_x in the model,

$$\begin{aligned} y_i^p = \mu + x_i\beta_x + c_i^T\beta_c + f_{\beta_{Age}}(Age_i) + Woman_i\beta_{Woman} \\ + Exsmoker_i\beta_{Exsmoker} + Smoker_i\beta_{Smoker} + CMV_i\beta_{CMV} \\ + PC1_i\beta_{PC1} + PC2_i\beta_{PC2} + SNP_i^T\beta_{SNP} + DayOfSampling_i + \varepsilon_i. \end{aligned} \qquad \textit{Eq. 3}$$

For age and sex, age and CMV serostatus, and age and smoking status, we also estimated their interaction effect by including one interaction term at a time in the model specified in *Eq. 3*. Hypothesis tests were performed by the Kenward-Roger approximation of the F-test for linear mixed models, implemented in the pbkrtest R package (Halekoh and Højsgaard, 2014).

The mediated effect

We estimated the mediated effect of aging, sex, variables related to smoking and CMV serostatus. It was estimated as the effect on 5mC levels mediated by changes in proportions of the 16 cell subsets due to the given factor. Estimates were computed by a two-stage procedure. Introduce the vector k_i of covariates: age (an entry for each spline term), sex, smoking, CMV serostatus and ancestry (2 PCs), but excluding the variable of interest, x (mediated effect of aging was estimated with a linear term), and let c_i be a vector of measured proportions of the 16 blood subsets. We fitted two different groups of models. In the first, measured proportions of immune cells were response variables. For the model of the n:th cell type, let β_k^n be the parameter vector for covariates k_i and β_x^n the parameter for the variable of interest. Let c_i^n denote the n:th entry of the vector c_i, the measured proportion of the n:th cell type for the i:th individual. For the model of 5mC levels in the M value scale at the p:th CpG site, y_i^p, let θ_x be a parameter for the variable of interest and θ_c and θ_k parameter vectors for the effects of cell proportions and covariates. To estimate mediated effects of the variable of interest x, we fit the models,

$$E\{c_i^n \mid x_i, k_i\} = \beta_0 + x_i\beta_x^n + k_i^T\beta_k^n, \text{ for } n = 1,..,16,$$

and

$$E\{y_i^p \mid x_i, c_i, k_i\} = \theta_0 + x_i\theta_x + c_i^T\theta_c + k_i^T\theta_k.$$

The mediated effect of x on DNA methylation was estimated by $\beta_x^T\theta_c$ (VanderWeele, 2015). Inference was done by the parametric bootstrap.

IDOL-adjusted effect

To compute the IDOL-adjusted effect, we estimated proportions of $CD4^+$ and $CD8^+$ T cells, B cells, NK cells, monocytes and neutrophils by the estimateCellCounts2 function in the FlowSorted.Blood.EPIC package with IDOL optimized CpG sites (Salas et al., 2018). For age, sex, smoking status and CMV serostatus, we estimated the IDOL-adjusted effect by adjusting for these estimated 6 proportions in the model specified by *Eq. 3*, instead of the 16 measured proportions.

Detection of the dispersion of DNA methylation with age

To estimate the change in dispersion of 5mC levels with age, we fit regression models where the residual variance depends on age. Let y_i^p be methylation levels on the M value scale for the p:th CpG site and the i:th individual. Using a similar notation as above, we estimated the dispersion effect of age by the parameter θ in the model,

$$y_i^p = \mu + x_i\beta_x + c_i^T\beta_c + SNP_i^T\beta_{SNP} + f_{\beta_{Age}}(Age_i) + Woman_i\beta_{Woman} + Exsmoker_i\beta_{Exsmoker} + Smoker_i\beta_{Smoker} + CMV_i\beta_{CMV} + PC1_i\beta_{PC1} + PC2_i\beta_{PC2} + \varepsilon_i, \quad \textit{Eq. 4}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\log\{\sigma\} = \tau + \text{Age}_i\theta$. We devised a hypothesis test for θ by a likelihood ratio test comparing that model to a model with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\log\{\sigma\} = \tau$ in *Eq. 4*. As a sensitivity analysis, we also fitted a model with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\log\{\sigma\} = \tau + c_i^T\beta_c + \text{Age}_i\theta$ in *Eq. 4*. Hypothesis test for θ in this case was done by comparing to a model with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\log\{\sigma\} = \tau + c_i^T\beta_c$ in *Eq. 4*. In this analysis, 77,708 CpG sites showed significant dispersion with age, 10% of which showed an increase in dispersion.

Cell-type specific methylation changes

Let y_i be 5mC levels measured at a CpG site on the β value scale for the i:th individual. Further, let m_i^l be 5mC levels and c_i^l measured proportions of the l:th cell type and x_i a vector of variables of interest. Expected average 5mC levels over all cells can be decomposed into,

$$E\{y_i \mid c_i^1, \ldots, c_i^L, x_i\} = \sum_l c_i^l E\{m_i^l \mid x_i\}. \quad \textit{Eq. 5}$$

Now assume that the expected value of m_i^l depends linearly on covariates of interest x_i,

$$E\{m_i^l \mid x_i\} = \mu^l + x_i^T\theta^l. \quad \textit{Eq. 6}$$

Inserting *Eq. 6* in *Eq. 5* yields

$$E\{y_i \mid c_i^1, \ldots, c_i^L, x_i\} = \sum_l c_i^l\mu^l + \sum_l c_i^l\, x_i^T\theta^l \quad \textit{Eq. 7}$$

Now, c_i^l are proportions, so

$$c_i^1 = 1 - \sum_{l \neq 1} c_i^l. \quad \textit{Eq. 8}$$

Inserting *Eq. 8* into *Eq. 7* and rearranging gives,

$$E\{y_i \mid c_i^1, \dots, c_i^L, x_i\} = \mu^1 + x_i^T\theta_1 + \sum_{l \neq 1} c_i^l(\mu^l - \mu^1) + \sum_{l \neq 1} c_i^l x_i^T(\theta^l - \theta^1), \qquad \textit{Eq. 9}$$

which, by a change of parameters $\mu = \mu^1$, $\beta = \theta_1$, $\tau^l = \mu^l - \mu^1$, $\beta^l = \theta^l - \theta$, gives the interaction model,

$$E\{y_i \mid c_i^1, \dots, c_i^L, x_i\} = \mu + x_i^T\beta + \sum_{l \neq 1} c_i^l(\tau^l) + \sum_{l \neq 1} c_i^l x_i^T\beta^l. \qquad \textit{Eq. 10}$$

We can now interpret the parameters of this model. The intercept μ corresponds to the intercept level of 5mC levels in a baseline cell type. The main effect terms β and τ^l are the effects of covariates x_i on 5mC levels in the baseline cell type and the difference in intercept term in the l:th cell type compared to the baseline cell type. Finally, the interaction parameter vector of interest β^l is the difference in the effect of the covariates on 5mC levels in the l:th cell type compared to the baseline cell type. Because it is the largest immune cell subset in blood, we used neutrophils as baseline cell type. This model had inflated variance for very small blood subsets. We therefore used the reduced set of 7 immune cells in this analysis. We estimated cell-specific effects on 5mC levels of age, sex and CMV serostatus by using,

$$x_i^T = (Age_i \quad Sex_i \quad CMV_i).$$

For cell-specific effects of genetic variants, we used a model that additionally included main effect terms for aging, sex and CMV serostatus,

$$E\{y_i \mid c_i^1, \dots, c_i^L, x_i\} = \mu + x_i^T\beta_x + k_i^T\beta_k + \sum_{l \neq 1} c_i^l(\tau^l) + \sum_{l \neq 1} c_i^l x_i^T\beta^l, \qquad \textit{Eq. 11}$$

where the vector k_i contains age (3 DoF spline term), sex and CMV serostatus, and $x_i = SNP_i^n$, the minor allele dosage of the n:th SNP associated with 5mC levels at the CpG site for the i:th individual. Inference was done by Wald tests with heteroscedasticity-consistent standard errors estimated by the sandwich R package (Zeileis et al., 2020). To test the interaction models in *Eq. 10* and *Eq. 11*, we performed simulations of a system like the one specified in *Eq. 5* and *Eq. 6*, except that the output of *Eq. 6* was logit-transformed to ensure that it was a proportion. We used observed

cell proportions for the simulations. When we simulated cell-type specific effects of aging in $CD8^+$ T cells and NK cells, drawn from a normal distribution with mean and standard deviation taken from those estimated for a moderate signal in the main effect age EWAS ($P_{adj} \approx 10^{-5}$), and zero effects of aging in the other cells, our model correctly detected cell-specific age effects in $CD8^+$ T cells and NK cells, but not in the other cells (Figure S3I).

Detection of gene × environment interactions

We tested whether age, sex, CMV serostatus, smoking status and CRP levels could have a different effect on the methylome depending on genotypes. For the i:th individual, let SNP_i^n be minor allele dosages of SNPs associated with 5mC levels at the p:th CpG site in the M value scale, y_i^p and let c_i be a vector of measured proportions of blood subsets with corresponding parameter vector β_c. Interaction effects for each variable of interest and each associated SNP were estimated for each CpG site in the model,

$$\begin{aligned} \mathrm{E}\{y_i^p \mid SNP_i^1, \ldots, SNP_i^N, Age_i, Woman_i, Smoker_i, CMV_i\} \\ = \mu + \sum_n SNP_i^n \beta_{SNP^n} + c_i^T \beta_c + PC1_i \beta_{PC1} + PC2_i \beta_{PC2} \\ + Age_i \beta_{Age} + Woman_i \beta_{Woman} + Smoker_i \beta_{Smoker} + CMV_i \beta_{CMV} \\ + \sum_n SNP_i^n \left(Age_i \theta_{Age}^n + Woman_i \theta_{Woman}^n + Smoker_i \theta_{Smoker}^n \right. \\ \left. + CMV_i \theta_{CMV}^n \right). \end{aligned} \qquad \textit{Eq. 12}$$

We investigated CRP in a separate model that simply added corresponding log-transformed CRP terms to *Eq. 12*. Inference was done by Wald tests with heteroscedasticity-consistent standard errors estimated by the sandwich R package (Zeileis et al., 2020).

Estimation of proportions of explained 5mC variance

According to our analyses, 5mC levels in the population are mainly associated with local genetic variation, blood cell composition, age, sex, smoking, CMV infection and CRP levels. We grouped these variables into 4 groups: genetic, cell composition, intrinsic and exposures. For a particular CpG site and the i:th individual, we collected observations of the minor allele dosage for the most associated local SNP in x_i^g, proportions of the 16 major cell types in the vector x_i^c, intrinsic factors (sex and natural spline expanded values of age) in the vector x_i^{in} and exposures (smoking status,

CMV serostatus and log-transformed CRP levels) in the vector x_e^i, with corresponding parameter vectors β_g, β_c, β_{in} and β_e. We interpret log-transformed CRP levels as a proxy measure of the exposure of chronic low-grade inflammation. For each group, we define the linear predictor terms:

$$f_g\left(x_i^g\right) = x_i^g \beta_g, \qquad \textit{Eq. 13}$$

$$f_c(x_i^c) = (x_i^c)^T \beta_c, \qquad \textit{Eq. 14}$$

$$f_{in}\left(x_i^{in}\right) = \left(x_i^{in}\right)^T \beta_{in}, \qquad \textit{Eq. 15}$$

$$f_e(x_i^e) = (x_i^e)^T \beta_e. \qquad \textit{Eq. 16}$$

These functions vary in complexity, so to get a fair comparison between them, we estimated group effect sizes as the out-of-sample proportion of variance explained by each group predictor. This estimation is done by indexing samples into two disjoint index groups I_1 and I_2, fitting the model on samples from I_1, and evaluating the prediction accuracy on samples from I_2.

Let y_i be 5mC levels at a CpG site on the M value scale. To compute the *total effect* of each group n on CpG methylation, we first fit the predictor function in individuals indexed to I_1,

$$\widehat{f_n}(x_i^n) = \hat{\mu} + (x_i^n)^T \widehat{\beta_n}, i \in I_1 \qquad \textit{Eq. 17}$$

with parameters estimated by least squares,

$$(\hat{\mu} \quad \widehat{\beta_n}) = \mathrm{argmin}_{\mu,\beta_n} \sum_{i \in I_1} \left(y_i - \mu - f_n(x_i^n)\right)^2. \qquad \textit{Eq. 18}$$

We can then define the *total effect size* for group n as the squared correlation between observations and the out-of-sample prediction,

$$(R_n^{\mathrm{Tot}})^2 = \mathrm{cor}\left(y_j, \widehat{f_n}\left(x_j^n\right)\right)^2, j \in I_2. \qquad \textit{Eq. 19}$$

For groups other than the cell composition group, we also computed a *direct effect*. For each group, it was computed as the added out-of-sample proportion of variance explained when adding the group predictor term to that of the cell composition group. The effect was computed for group n by

$$(R_n^D)^2 = (R_{n+c}^{Tot})^2 - (R_c^{Tot})^2 \qquad \textit{Eq. 20}$$

Where $(R_{n+c}^{\mathrm{Tot}})^2$ is the total effect of the predictor including both group n and proportions of cell types

$$\widehat{f_{c+n}}(x_i^n, x_i^c) = \hat{\mu} + (x^n)_i^T \widehat{\beta_n} + (x^c)_i^T \widehat{\beta_c}. \qquad \textit{Eq. 21}$$

To mitigate the impact of sampling on estimates of *total* and *direct effects*, we did four independent repeats of fivefold cross-validation and averaged the effect sizes across all 20 drawn samples. To have an unbiased estimation of the out-of-sample explained variance, we redid a local meQTL mapping on the training set in each iteration of the cross-validation scheme. The algorithm for drawing samples of the total effect is detailed in *Algorithm 2*.

Algorithm 2: Cross-validation for estimating out-of-sample group total effect size

---

Repeat 4 times:

For $k = 1, \ldots, 5$

Index a fifth of individuals as I_k, the others are indexed as $I_{\setminus k}$

Select SNP for the predictor f_g by performing a local meQTL mapping on individuals in $I_{\setminus k}$

For predictor $f_n \in \{f_g, f_c, f_{in}, f_e\}$

Estimate $\widehat{f_n}$ by *Eq. 17* and *Eq. 18* with $I_1 = I_{\setminus k}$

Compute $(R_n^{\mathrm{Tot}})^2$ by *Eq. 19* with $I_2 = I_k$

Finally, we computed an effect size for interactions between genetic and non-genetic factors. It was computed, similar to *Eq. 20*, as the added out-of-sample proportion of variance explained by the regression function,

$$f_{Int}(Age_i, Woman_i, CMV_i, ExSmoker_i, Smoker_i, CRP_i) \qquad \textit{Eq. 22}$$
$$= \mu + SNP_i\beta_{SNP} + Age_i\beta_{Age} + Woman_i\beta_{Woman} + CMV_i\beta_{CMV}$$
$$+ ExSmoker_i\beta_{ExSmoker} + Smoker_i\beta_{Smoker} + log(CRP_i)\beta_{CRP}$$
$$+ SNP_i\left(Age_i\beta_{Age}^{SNP} + Woman_i\beta_{Woman}^{SNP} + CMV_i\beta_{CMV}^{SNP}\right.$$
$$\left.+ ExSmoker_i\beta_{ExSmoker}^{SNP} + Smoker_i\beta_{Smoker}^{SNP} + log(CRP_i)\beta_{CRP}^{SNP}\right),$$

compared to the same regression function without interaction terms:

$$f_{Main}(Age_i, Woman_i, CMV_i, ExSmoker_i, Smoker_i, CRP_i) \qquad \textit{Eq. 23}$$
$$= \mu + SNP_i\beta_{SNP} + Age_i\beta_{Age} + Woman_i\beta_{Woman} + CMV_i\beta_{CMV}$$
$$+ ExSmoker_i\beta_{ExSmoker} + Smoker_i\beta_{Smoker} + log(CRP_i)\beta_{CRP}.$$

Biological annotations

Information about the position, closest gene and CpG density of each CpG site was obtained from the Illumina EPIC array manifest v.1.0 B4. We retrieved the chromatin state of regions around each CpG site, using the 15 chromatin states inferred with ChromHMM for $CD4^+$ naive T cells by the ROADMAP Epigenomics consortium (Roadmap Epigenomics et al., 2015). We used $CD4^+$ naive T cells as a reference because it is a large, relatively homogeneous subset of cells that are less differentiated than memory cells. We obtained similar results when using other cell subsets as reference (data not shown). The data was downloaded from the consortium webpage (https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html). The transcription factor binding site data used was public CHIP-seq data collected and processed for the 2020 release of the ReMap database (Cheneby et al., 2020), including a total of 1,165 TFs. Binding sites include both direct and indirect binding. Enrichment analyses were performed by creating a simple two-way table for each target set and TF, and then performing a Fisher's exact test. Gene ontology enrichments were computed with the gometh function in the missMethyl R package (Phipson et al., 2016).

We tested if a set of x local or remote meQTL SNPs is enriched in disease- or trait-associated variants, by sampling at random, among all tested SNPs, 15,000 sets of x SNPs with minor allele frequencies matched to those of meQTL SNPs. For each resampled set, we calculated the proportion of variants either known to be associated with a disease or trait, or in linkage disequilibrium (LD; set here to $r^2 > 0.6$) with a disease/trait-associated variant (P-value $< 5\times10^{-8}$;

EBI-NHGRI Catalog of GWAS hits version e100 r2021-01-1). The enrichment *P*-value was estimated as the percentage of resamples for which this proportion was larger than that observed in meQTL SNPs. LD was precomputed for all 5,699,237 SNPs with PLINK 1.9 (with arguments '–show-tags all–tag-kb 500–tag-r2 0.6') (Chang et al., 2015).

● DATA AND SOFTWARE AVAILABILITY

The Infinium MethylationEPIC raw intensity data have been deposited at the European Genome-phenome Archive (EGA, https://www.ebi.ac.uk/ega/) under accession number EGAS0000XXXXXX. Data access applications are reviewed by a data access committee (DAC) and access is granted if the request is consistent with the consent provided by *Milieu Intérieur* participants. All association statistics obtained in this study (i.e., local meQTL mapping, the 141 EWAS and interaction models) can be explored and downloaded from the web browser hub05.hosting.pasteur.fr/MIMETH_browser/. All the code supporting the current study has been uploaded to GitHub: https://github.com/JacobBergstedt/mimeth.

SUPPLEMENTAL INFORMATION

The supplemental PDF includes 6 supplemental figures.

Supplemental tables are provided as 6 separate excel files:

Table S1. Candidate intrinsic and environmental factors tested for association with the blood DNA methylome of adults.

Table S2. Summary statistics for significant remote-effect meQTLs.

Table S3. Significant enrichments of variable-associated CpG sites in binding sites for transcription factors (TFs).

Table S4. Significant gene ontology enrichments of genes close to variable-associated CpG sites.

Table S5. CpG sites significantly associated with two interacting variables.

Table S6. Proportions of variance explained by intrinsic factors, exposures, cell composition and local SNPs for the 10,000 CpG sites with the most explained variance.

REFERENCES

Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., and Abecasis, G.R. (2015). A global reference for human genetic variation. Nature *526*, 68-74.

Beerman, I., Bock, C., Garrison, B.S., Smith, Z.D., Gu, H., Meissner, A., and Rossi, D.J. (2013). Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. Cell Stem Cell *12*, 413-425.

Bell, J.T., Tsai, P.C., Yang, T.P., Pidsley, R., Nisbet, J., Glass, D., Mangino, M., Zhai, G., Zhang, F., Valdes, A., *et al.* (2012). Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet *8*, e1002629.

Berdasco, M., and Esteller, M. (2019). Clinical epigenetics: seizing opportunities for translation. Nat Rev Genet *20*, 109-127.

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., *et al.* (2017). Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet *49*, 131-138.

Boyce, W.T., and Kobor, M.S. (2015). Development and the epigenome: the 'synapse' of gene-environment interplay. Dev Sci *18*, 1-23.

Boyer, L.A., Plath, K., Zeitlinger, J., Brambrink, T., Medeiros, L.A., Lee, T.I., Levine, S.S., Wernig, M., Tajonar, A., Ray, M.K., *et al.* (2006). Polycomb complexes repress developmental regulators in murine embryonic stem cells. Nature *441*, 349-353.

Bracken, A.P., Kleine-Kohlbrecher, D., Dietrich, N., Pasini, D., Gargiulo, G., Beekman, C., Theilgaard-Monch, K., Minucci, S., Porse, B.T., Marine, J.C., *et al.* (2007). The Polycomb group proteins bind throughout the INK4A-ARF locus and are disassociated in senescent cells. Genes Dev *21*, 525-530.

Braun, K.V.E., Dhana, K., de Vries, P.S., Voortman, T., van Meurs, J.B.J., Uitterlinden, A.G., consortium, B., Hofman, A., Hu, F.B., Franco, O.H., *et al.* (2017). Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. Clin Epigenetics *9*, 15.

Bush, N.R., Edgar, R.D., Park, M., MacIsaac, J.L., McEwen, L.M., Adler, N.E., Essex, M.J., Kobor, M.S., and Boyce, W.T. (2018). The biological embedding of early-life socioeconomic status and family adversity in children's genome-wide DNA methylation. Epigenomics *10*, 1445-1461.

Cannon, M.J., Schmid, D.S., and Hyde, T.B. (2010). Review of cytomegalovirus seroprevalence and demographic characteristics associated with infection. Rev Med Virol *20*, 202-213.

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

Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience *4*, 7.

Cheneby, J., Menetrier, Z., Mestdagh, M., Rosnet, T., Douida, A., Rhalloussi, W., Bergon, A., Lopez, F., and Ballester, B. (2020). ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. Nucleic Acids Res *48*, D180-D188.

Chong, S.J.F., Marchi, S., Petroni, G., Kroemer, G., Galluzzi, L., and Pervaiz, S. (2020). Noncanonical Cell Fate Regulation by Bcl-2 Proteins. Trends Cell Biol *30*, 537-555.

Correa-Saez, A., Jimenez-Izquierdo, R., Garrido-Rodriguez, M., Morrugares, R., Munoz, E., and Calzado, M.A. (2020). Updating dual-specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2): molecular basis, functions and role in diseases. Cell Mol Life Sci *77*, 4747-4763.

Crinier, A., Milpied, P., Escaliere, B., Piperoglou, C., Galluso, J., Balsamo, A., Spinelli, L., Cervera-Marzal, I., Ebbo, M., Girard-Madoux, M., *et al.* (2018). High-Dimensional Single-Cell Analysis Identifies Organ-Specific Signatures and Conserved NK Cell Subsets in Humans and Mice. Immunity *49*, 971-986 e975.

Czamara, D., Eraslan, G., Page, C.M., Lahti, J., Lahti-Pulkkinen, M., Hamalainen, E., Kajantie, E., Laivuori, H., Villa, P.M., Reynolds, R.M., *et al.* (2019). Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nat Commun *10*, 2548.

Dozmorov, M.G. (2015). Polycomb repressive complex 2 epigenomic signature defines age-associated hypermethylation and gene expression changes. Epigenetics *10*, 484-495.

Dugue, P.A., Jung, C.H., Joo, J.E., Wang, X., Wong, E.M., Makalic, E., Schmidt, D.F., Baglietto, L., Severi, G., Southey, M.C., *et al.* (2020). Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. Epigenetics *15*, 358-368.

Ek, W.E., Ahsan, M., Rask-Andersen, M., Liang, L., Moffatt, M.F., Gyllensten, U., and Johansson, A. (2017). Epigenome-wide DNA methylation study of IgE concentration in relation to self-reported allergies. Epigenomics *9*, 407-418.

Farlik, M., Halbritter, F., Muller, F., Choudry, F.A., Ebert, P., Klughammer, J., Farrow, S., Santoro, A., Ciaurro, V., Mathur, A., *et al.* (2016). DNA Methylation Dynamics of Human Hematopoietic Stem Cell Differentiation. Cell Stem Cell *19*, 808-822.

Feil, R., and Fraga, M.F. (2012). Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet *13*, 97-109.

Feinberg, A.P. (2018). The Key Role of Epigenetics in Human Disease Prevention and Mitigation. N Engl J Med *378*, 1323-1334.

Fleiss, J.L. (2011). Design and analysis of clinical experiments, Vol 73 (New York: Wiley).

Fortin, J.P., Triche, T.J., Jr., and Hansen, K.D. (2017). Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. Bioinformatics *33*, 558-560.

Fraga, M.F., Ballestar, E., Paz, M.F., Ropero, S., Setien, F., Ballestar, M.L., Heine-Suner, D., Cigudosa, J.C., Urioste, M., Benitez, J., *et al.* (2005). Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A *102*, 10604-10609.

Furman, D., Campisi, J., Verdin, E., Carrera-Bastos, P., Targ, S., Franceschi, C., Ferrucci, L., Gilroy, D.W., Fasano, A., Miller, G.W., *et al.* (2019). Chronic inflammation in the etiology of disease across the life span. Nat Med *25*, 1822-1832.

Gao, X., Jia, M., Zhang, Y., Breitling, L.P., and Brenner, H. (2015). DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. Clin Epigenetics *7*, 113.

Garvin, A.J., Densham, R.M., Blair-Reid, S.A., Pratt, K.M., Stone, H.R., Weekes, D., Lawrence, K.J., and Morris, J.R. (2013). The deSUMOylase SENP7 promotes chromatin relaxation for homologous recombination DNA repair. EMBO Rep *14*, 975-983.

Goronzy, J.J., Hu, B., Kim, C., Jadhav, R.R., and Weyand, C.M. (2018). Epigenetics of T cell aging. J Leukoc Biol *104*, 691-699.

Groves, I.J., Jackson, S.E., Poole, E.L., Nachshon, A., Rozman, B., Schwartz, M., Prinjha, R.K., Tough, D.F., Sinclair, J.H., and Wills, M.R. (2021). Bromodomain proteins regulate human cytomegalovirus latency and reactivation allowing epigenetic therapeutic intervention. Proc Natl Acad Sci U S A *118*.

Grundberg, E., Small, K.S., Hedman, A.K., Nica, A.C., Buil, A., Keildson, S., Bell, J.T., Yang, T.P., Meduri, E., Barrett, A., *et al.* (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet *44*, 1084-1089.

Halekoh, U., and Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. J Stat Soft *59*, 32.

Hannon, E., Gorrie-Stone, T.J., Smart, M.C., Burrage, J., Hughes, A., Bao, Y., Kumari, M., Schalkwyk, L.C., and Mill, J. (2018). Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. Am J Hum Genet *103*, 654-665.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.B., Gao, Y., *et al.* (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell *49*, 359-367.

Heyn, H., Li, N., Ferreira, H.J., Moran, S., Pisano, D.G., Gomez, A., Diez, J., Sanchez-Mut, J.V., Setien, F., Carmona, F.J., *et al.* (2012). Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A *109*, 10522-10527.

Holler, C.J., Webb, R.L., Laux, A.L., Beckett, T.L., Niedowicz, D.M., Ahmed, R.R., Liu, Y., Simmons, C.R., Dowling, A.L., Spinelli, A., *et al.* (2012). BACE2 expression increases in human neurodegenerative disease. Am J Pathol *180*, 337-350.

Hop, P.J., Luijk, R., Daxinger, L., van Iterson, M., Dekkers, K.F., Jansen, R., Consortium, B., van Meurs, J.B.J., t Hoen, P.A.C., Ikram, M.A., *et al.* (2020). Genome-wide identification of genes regulating DNA methylation using genetic anchors for causal inference. Genome Biol *21*, 220.

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

Houseman, E.A., Kelsey, K.T., Wiencke, J.K., and Marsit, C.J. (2015). Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. BMC Bioinformatics *16*, 95.

Hwang, J.Y., Aromolaran, K.A., and Zukin, R.S. (2017). The emerging field of epigenetics in neurodegeneration and neuroprotection. Nat Rev Neurosci *18*, 347-361.

Jaffe, A.E., and Irizarry, R.A. (2014). Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol *15*, R31.

Johansson, A., Enroth, S., and Gyllensten, U. (2013). Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. PLoS One *8*, e67378.

Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics *8*, 118-127.

Jones, M.J., Goodman, S.J., and Kobor, M.S. (2015). DNA methylation and healthy human aging. Aging Cell *14*, 924-932.

Jost, E., Lin, Q., Weidner, C.I., Wilop, S., Hoffmann, M., Walenda, T., Schemionek, M., Herrmann, O., Zenke, M., Brummendorf, T.H., *et al.* (2014). Epimutations mimic genomic mutations of DNMT3A in acute myeloid leukemia. Leukemia *28*, 1227-1234.

Karlsson Linner, R., Biroli, P., Kong, E., Meddens, S.F.W., Wedow, R., Fontana, M.A., Lebreton, M., Tino, S.P., Abdellaoui, A., Hammerschlag, A.R., *et al.* (2019). Genome-wide association analyses of risk tolerance and risky behaviors in over 1 million individuals identify hundreds of loci and shared genetic influences. Nat Genet *51*, 245-257.

Karlsson Linner, R., Marioni, R.E., Rietveld, C.A., Simpkin, A.J., Davies, N.M., Watanabe, K., Armstrong, N.J., Auro, K., Baumbach, C., Bonder, M.J., *et al.* (2017). An epigenome-wide association study meta-analysis of educational attainment. Mol Psychiatry *22*, 1680-1690.

Kim, J.Y., Tavare, S., and Shibata, D. (2005). Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. Proc Natl Acad Sci U S A *102*, 17739-17744.

Klenerman, P., and Oxenius, A. (2016). T cell responses to cytomegalovirus. Nat Rev Immunol *16*, 367-377.

Koestler, D.C., Jones, M.J., Usset, J., Christensen, B.C., Butler, R.A., Kobor, M.S., Wiencke, J.K., and Kelsey, K.T. (2016). Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics *17*, 120.

Lam, L.L., Emberly, E., Fraser, H.B., Neumann, S.M., Chen, E., Miller, G.E., and Kobor, M.S. (2012). Factors underlying variable DNA methylation in a human community cohort. Proc Natl Acad Sci U S A *109 Suppl 2*, 17253-17260.

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

Lee, J.J., Wedow, R., Okbay, A., Kong, E., Maghzian, O., Zacher, M., Nguyen-Viet, T.A., Bowers, P., Sidorenko, J., Karlsson Linner, R., *et al.* (2018). Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat Genet *50*, 1112-1121.

Lemire, M., Zaidi, S.H., Ban, M., Ge, B., Aissi, D., Germain, M., Kassam, I., Wang, M., Zanke, B.W., Gagnon, F., *et al.* (2015). Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat Commun *6*, 6326.

Li, C., Guo, R., Lou, J., and Zhou, H. (2012). The transcription levels of ABCA1, ABCG1 and SR-BI are negatively associated with plasma CRP in Chinese populations with various risk factors for atherosclerosis. Inflammation *35*, 1641-1648.

Li, S., Wong, E.M., Nguyen, T.L., Joo, J.E., Stone, J., Dite, G.S., Giles, G.G., Saffery, R., Southey, M.C., and Hopper, J.L. (2017). Causes of blood methylomic variation for middle-aged women measured by the HumanMethylation450 array. Epigenetics *12*, 973-981.

Li, Y., Zheng, H., Wang, Q., Zhou, C., Wei, L., Liu, X., Zhang, W., Zhang, Y., Du, Z., Wang, X., *et al.* (2018). Genome-wide analyses reveal a role of Polycomb in promoting hypomethylation of DNA methylation valleys. Genome Biol *19*, 18.

Liang, L., Willis-Owen, S.A.G., Laprise, C., Wong, K.C.C., Davies, G.A., Hudson, T.J., Binia, A., Hopkin, J.M., Yang, I.V., Grundberg, E., *et al.* (2015). An epigenome-wide association study of total serum immunoglobulin E concentration. Nature *520*, 670-674.

Ligthart, S., Marzi, C., Aslibekyan, S., Mendelson, M.M., Conneely, K.N., Tanaka, T., Colicino, E., Waite, L.L., Joehanes, R., Guan, W., *et al.* (2016). DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. Genome Biol *17*, 255.

Ligthart, S., Vaez, A., Vosa, U., Stathopoulou, M.G., de Vries, P.S., Prins, B.P., Van der Most, P.J., Tanaka, T., Naderi, E., Rose, L.M., *et al.* (2018). Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. Am J Hum Genet *103*, 691-706.

Ling, C., and Ronn, T. (2019). Epigenetics in Human Obesity and Type 2 Diabetes. Cell Metab *29*, 1028-1044.

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

Manichaikul, A., Mychaleckyj, J.C., Rich, S.S., Daly, K., Sale, M., and Chen, W.M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics *26*, 2867-2873.

Marchal, C., and Miotto, B. (2015). Emerging concept in DNA methylation: role of transcription factors in shaping DNA methylation patterns. J Cell Physiol *230*, 743-751.

Martin, E.M., and Fry, R.C. (2018). Environmental Influences on the Epigenome: Exposure-Associated DNA Methylation in Human Populations. Annu Rev Public Health *39*, 309-333.

Mason, G.M., Poole, E., Sissons, J.G., Wills, M.R., and Sinclair, J.H. (2012). Human cytomegalovirus latency alters the cellular secretome, inducing cluster of differentiation (CD)4+ T-cell migration and suppression of effector function. Proc Natl Acad Sci U S A *109*, 14538-14543.

Mauvais-Jarvis, F., Bairey Merz, N., Barnes, P.J., Brinton, R.D., Carrero, J.J., DeMeo, D.L., De Vries, G.J., Epperson, C.N., Govindan, R., Klein, S.L., *et al.* (2020). Sex and gender: modifiers of health, disease, and medicine. Lancet *396*, 565-582.

Mazzone, R., Zwergel, C., Artico, M., Taurone, S., Ralli, M., Greco, A., and Mai, A. (2019). The emerging role of epigenetics in human autoimmune disorders. Clin Epigenetics *11*, 34.

McCartney, D.L., Zhang, F., Hillary, R.F., Zhang, Q., Stevenson, A.J., Walker, R.M., Bermingham, M.L., Boutin, T., Morris, S.W., Campbell, A., *et al.* (2019). An epigenome-wide association study of sex-specific chronological ageing. Genome Med *12*, 1.

Michalak, E.M., Burr, M.L., Bannister, A.J., and Dawson, M.A. (2019). The roles of DNA, RNA and histone methylation in ageing and cancer. Nat Rev Mol Cell Biol *20*, 573-589.

Morita, K., Okamura, T., Inoue, M., Komai, T., Teruya, S., Iwasaki, Y., Sumitomo, S., Shoda, H., Yamamoto, K., and Fujio, K. (2016). Egr2 and Egr3 in regulatory T cells cooperatively control systemic autoimmunity through Ltbp3-mediated TGF-beta3 production. Proc Natl Acad Sci U S A *113*, E8131-E8140.

Niccoli, T., and Partridge, L. (2012). Ageing as a risk factor for disease. Curr Biol *22*, R741-752.

Nikolich-Zugich, J. (2018). The twilight of immunity: emerging concepts in aging of the immune system. Nat Immunol *19*, 10-19.

O'Geen, H., Squazzo, S.L., Iyengar, S., Blahnik, K., Rinn, J.L., Chang, H.Y., Green, R., and Farnham, P.J. (2007). Genome-wide analysis of KAP1 binding suggests autoregulation of KRAB-ZNFs. PLoS Genet *3*, e89.

Patin, E., Hasan, M., Bergstedt, J., Rouilly, V., Libri, V., Urrutia, A., Alanio, C., Scepanovic, P., Hammer, C., Jonsson, F., *et al.* (2018). Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. Nat Immunol *19*, 302-314.

Perumal, N., Funke, S., Pfeiffer, N., and Grus, F.H. (2016). Proteomics analysis of human tears from aqueous-deficient and evaporative dry eye patients. Sci Rep *6*, 29629.

Peterson, C.B., Bogomolov, M., Benjamini, Y., and Sabatti, C. (2016). Many Phenotypes Without Many False Discoveries: Error Controlling Strategies for Multitrait Association Studies. Genet Epidemiol *40*, 45-56.

Phipson, B., Maksimovic, J., and Oshlack, A. (2016). missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics *32*, 286-288.

Price, M.E., Cotton, A.M., Lam, L.L., Farre, P., Emberly, E., Brown, C.J., Robinson, W.P., and Kobor, M.S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin *6*, 4.

Quenneville, S., Turelli, P., Bojkowska, K., Raclot, C., Offner, S., Kapopoulou, A., and Trono, D. (2012). The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. Cell Rep *2*, 766-773.

Roadmap Epigenomics, C., Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., *et al.* (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317-330.

Salas, L.A., Koestler, D.C., Butler, R.A., Hansen, H.M., Wiencke, J.K., Kelsey, K.T., and Christensen, B.C. (2018). An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol *19*, 64.

Samet, J.M. (2013). Tobacco smoking: the leading cause of preventable disease worldwide. Thorac Surg Clin *23*, 103-112.

Savva, G.M., Pachnio, A., Kaul, B., Morgan, K., Huppert, F.A., Brayne, C., Moss, P.A., Medical Research Council Cognitive, F., and Ageing, S. (2013). Cytomegalovirus infection is associated with increased mortality in the older population. Aging Cell *12*, 381-387.

Shabalin, A.A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics *28*, 1353-1358.

Siebold, A.P., Banerjee, R., Tie, F., Kiss, D.L., Moskowitz, J., and Harte, P.J. (2010). Polycomb Repressive Complex 2 and Trithorax modulate Drosophila longevity and stress resistance. Proc Natl Acad Sci U S A *107*, 169-174.

Singmann, P., Shem-Tov, D., Wahl, S., Grallert, H., Fiorito, G., Shin, S.Y., Schramm, K., Wolf, P., Kunze, S., Baran, Y., *et al.* (2015). Characterization of whole-genome autosomal differences of DNA methylation between men and women. Epigenetics Chromatin *8*, 43.

Sproston, N.R., and Ashworth, J.J. (2018). Role of C-Reactive Protein at Sites of Inflammation and Infection. Front Immunol *9*, 754.

Stekhoven, D.J., and Buhlmann, P. (2012). MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics *28*, 112-118.

Stringhini, S., Polidoro, S., Sacerdote, C., Kelly, R.S., van Veldhoven, K., Agnoli, C., Grioni, S., Tumino, R., Giurdanella, M.C., Panico, S., *et al.* (2015). Life-course socioeconomic status and DNA methylation of genes regulating inflammation. Int J Epidemiol *44*, 1320-1330.

Teh, A.L., Pan, H., Chen, L., Ong, M.L., Dogra, S., Wong, J., MacIsaac, J.L., Mah, S.M., McEwen, L.M., Saw, S.M., *et al.* (2014). The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res *24*, 1064-1074.

Teschendorff, A.E., Breeze, C.E., Zheng, S.C., and Beck, S. (2017). A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinformatics *18*, 105.

Teschendorff, A.E., Menon, U., Gentry-Maharaj, A., Ramus, S.J., Weisenberger, D.J., Shen, H., Campan, M., Noushmehr, H., Bell, C.G., Maxwell, A.P., *et al.* (2010). Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res *20*, 440-446.

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

Thomas, S., Rouilly, V., Patin, E., Alanio, C., Dubois, A., Delval, C., Marquier, L.G., Fauchoux, N., Sayegrih, S., Vray, M., *et al.* (2015). The Milieu Interieur study - an integrative approach for study of human immunological variance. Clin Immunol *157*, 277-293.

Timp, W., and Feinberg, A.P. (2013). Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. Nat Rev Cancer *13*, 497-510.

Torti, N., Walton, S.M., Murphy, K.M., and Oxenius, A. (2011). Batf3 transcription factor-dependent DC subsets in murine CMV infection: differential impact on T-cell priming and memory inflation. Eur J Immunol *41*, 2612-2618.

Triche, T.J., Jr., Weisenberger, D.J., Van Den Berg, D., Laird, P.W., and Siegmund, K.D. (2013). Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res *41*, e90.

Tserel, L., Kolde, R., Limbach, M., Tretyakov, K., Kasela, S., Kisand, K., Saare, M., Vilo, J., Metspalu, A., Milani, L., *et al.* (2015). Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. Sci Rep *5*, 13107.

van der Harst, P., de Windt, L.J., and Chambers, J.C. (2017). Translational Perspective on Epigenetics in Cardiovascular Disease. J Am Coll Cardiol *70*, 590-606.

van Dongen, J., Nivard, M.G., Willemsen, G., Hottenga, J.J., Helmer, Q., Dolan, C.V., Ehli, E.A., Davies, G.E., van Iterson, M., Breeze, C.E., *et al.* (2016). Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun *7*, 11115.

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

Villicana, S., and Bell, J.T. (2021). Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol *22*, 127.

Vogel, M.J., Guelen, L., de Wit, E., Peric-Hupkes, D., Loden, M., Talhout, W., Feenstra, M., Abbas, B., Classen, A.K., and van Steensel, B. (2006). Human heterochromatin proteins form large domains containing KRAB-ZNF genes. Genome Res *16*, 1493-1504.

Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Kasela, S., *et al.* (2018). Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. bioRxiv.

Wang, Y., Karlsson, R., Lampa, E., Zhang, Q., Hedman, A.K., Almgren, M., Almqvist, C., McRae, A.F., Marioni, R.E., Ingelsson, E., *et al.* (2018). Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. Epigenetics *13*, 975-987.

Wild, C.P. (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev *14*, 1847-1850.

Williams, K., Christensen, J., and Helin, K. (2011). DNA methylation: TET proteins-guardians of CpG islands? EMBO Rep *13*, 28-35.

Xu, L., Li, X., Chu, E.S., Zhao, G., Go, M.Y., Tao, Q., Jin, H., Zeng, Z., Sung, J.J., and Yu, J. (2012a). Epigenetic inactivation of BCL6B, a novel functional tumour suppressor for gastric cancer, is associated with poor survival. Gut *61*, 977-985.

Xu, Y., Tarquini, F., Romero, R., Kim, C.J., Tarca, A.L., Bhatti, G., Lee, J., Sundell, I.B., Mittal, P., Kusanovic, J.P., *et al.* (2012b). Peripheral CD300a+CD8+ T lymphocytes with a distinct cytotoxic molecular signature increase in pregnant women with chronic chorioamnionitis. Am J Reprod Immunol *67*, 184-197.

Yang, Z., Wong, A., Kuh, D., Paul, D.S., Rakyan, V.K., Leslie, R.D., Zheng, S.C., Widschwendter, M., Beck, S., and Teschendorff, A.E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol *17*, 205.

Yousefi, P., Huen, K., Dave, V., Barcellos, L., Eskenazi, B., and Holland, N. (2015). Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. BMC Genomics *16*, 911.

Yusipov, I., Bacalini, M.G., Kalyakulina, A., Krivonosov, M., Pirazzini, C., Gensous, N., Ravaioli, F., Milazzo, M., Giuliani, C., Vedunova, M., *et al.* (2020). Age-related DNA methylation changes are sex-specific: a comprehensive assessment. Aging (Albany NY) *12*, 24057-24080.

Zeileis, A., Köll, S., and Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in R. J Stat Soft *95*, 36.

Zhao, M., Kim, P., Mitra, R., Zhao, J., and Zhao, Z. (2016). TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. Nucleic Acids Res *44*, D1023-1031.

Zheng, S.C., Breeze, C.E., Beck, S., and Teschendorff, A.E. (2018). Identification of differentially methylated cell types in epigenome-wide association studies. Nat Methods *15*, 1059-1066.

Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., and Berman, B.P. (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet *50*, 591-602.

Zuo, X., Sheng, J., Lau, H.T., McDonald, C.M., Andrade, M., Cullen, D.E., Bell, F.T., Iacovino, M., Kyba, M., Xu, G., *et al.* (2012). Zinc finger protein ZFP57 requires its co-factor to recruit DNA methyltransferases and maintains DNA methylation imprint in embryonic stem cells via its transcriptional repression domain. J Biol Chem *287*, 2107-2118.