The immune factors driving DNA methylation variation in human blood
Jacob Bergstedt, Sadoune Ait Kaci Azzou, Kristin Tsuo, Anthony Jaquaniello, Alejandra Urrutia, Maxime Rotival, David Lin, Julia Macisaac, Michael Kobor, Matthew Albert, et al.

To cite this version:
Jacob Bergstedt, Sadoune Ait Kaci Azzou, Kristin Tsuo, Anthony Jaquaniello, Alejandra Urrutia, et al.. The immune factors driving DNA methylation variation in human blood. Nature Communications, 2022, 13 (1), pp.5895. 10.1038/s41467-022-33511-6. pasteur-03865045

HAL Id: pasteur-03865045
https://pasteur.hal.science/pasteur-03865045v1
Submitted on 22 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License
The immune factors driving DNA methylation variation in human blood

Jacob Bergstedt, Sadoune Ait Kaci Azzou, Kristin Tsuo, Anthony Jaquaniello, Alejandra Urrutia, Maxime Rotival, David T. S. Lin, Julia L. Maclsaac, Michael S. Kobor, Matthew L. Albert, Darragh Duffy, Etienne Patin, Lluís Quintana-Murci, & Milieu Intérieur Consortium*

Epigenetic changes are required for normal development, 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 884 adults is affected by DNA sequence variation, age, sex and 139 factors relating to life habits and immunity. Furthermore, we investigated whether these effects are mediated or not by changes in cellular composition, measured by deep immunophenotyping. We show that DNA methylation differs substantially between naïve and memory T cells, supporting the need for adjustment on these cell-types. By doing so, we find that latent cytomegalovirus infection drives DNA methylation variation and provide further support that the increased dispersion of DNA methylation with aging is due to epigenetic drift. Finally, our results indicate that cellular composition and DNA sequence variation are the strongest predictors of DNA methylation, highlighting critical factors for medical epigenomics studies.

Epigenetic research has improved our understanding of the existing links between environmental risk factors, aging, genetic variation, and human disease. Epigenome-wide association studies (EWAS) have shown that DNA methylation (i.e., 5-methylcytosine, 5mC), the most studied epigenetic mark in humans, is associated with a wide range of environmental exposures along the life course, such as chemicals or past socioeconomic status. 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. These studies collectively suggest that DNA methylation marks could be of tremendous value as gauges of the exposome and as clinical biomarkers.

However, interpretation of EWAS remains limited. First, because the epigenome of a cell reflects its identity, a risk factor or a disease that alters cellular composition also alters 5mC levels measured in the tissue. 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. Previous studies have accounted for cellular heterogeneity in blood by using cell sorting experiments, or cellular proportions estimated from 5mC profiles through in-silico cell mixture deconvolution.
but these approaches focus on a subset of frequent cell types that capture only a part of blood cellular composition. Second, the strong links between DNA methylation and DNA sequence variation, attested by the numerous DNA methylation quantitative trait loci (meQTLs) detected so far, suggest that environmental effects on the epigenome may operate through gene-by-environment interactions, but evidence for such interactions remains currently minimal. Finally, environmental risk factors with a yet-unidentified effect on DNA methylation, such as common infections, could 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 884 healthy adults of the Milieu Intérieur cohort. We leveraged the deep characterization of the cohort, including high-resolution immunophenotyping by flow cytometry, to determine whether and how cellular composition, intrinsic factors (i.e., age and sex), genetic variation, and 139 health and immunity-related variables and environmental exposures affect the blood DNA methylome. We first assessed differences in the DNA methylation profiles of 16 different immune cell types. We then performed EWAS, adjusted or not for the measured proportions of the 16 immune cell subsets, and mediation analyses 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.

We show that adjusting EWAS for 16 measured cell proportions better accounts for cellular heterogeneity than current cell mixture deconvolution methods. We identify latent cell types, such as CD8+ T cells, as key factors affecting population variation in SmC levels, through the dysregulation of human transcription factors and profound changes in the proportion of differentiated T cells. We show that the increased dispersion of DNA methylation with aging is independent of cellular composition, supporting instead a decrease in the fidelity of the epigenetic maintenance machinery. Furthermore, 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 subtle changes in blood cell composition, and characterize the DNA methylation signature of cell types affected by these factors. Finally, 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. 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

#### Proportions of naïve and differentiated T cells markedly contribute to DNA methylation variation

To investigate the non-genetic and genetic factors that affect population variation in DNA methylation, we quantified SmC levels at >850,000 CpG sites, with the Illumina Infinium MethylationEPIC array, in the 1000 healthy donors of the Milieu Intérieur cohort (Fig. 1a). 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 factors that are known to affect DNA methylation (i.e., age, sex, smoking, BMI and socioeconomic status), that have been proposed to affect DNA methylation (e.g., dietary habits, upbringing) or that pertain to the immune system (e.g., past and latent infections, past vaccinations, antibody levels; Supplementary Data 1). All donors were genotyped at 945,213 single-nucleotide polymorphisms (SNPs), yielding 5,699,237 accurate SNPs after imputation. After quality control filtering, high-quality measurements of DNA methylation were obtained at 644,517 CpG sites for 884 unrelated individuals (Supplementary Fig. 1; Methods). We found that SmC levels well reproduce expected patterns across chromatin states, supporting the good quality of the data (Supplementary Fig. 1 and Supplementary Notes). Whereas most epigenome-wide studies adopt an estimated cellular composition to detect direct effects on DNA methylation (i.e., acting through changes within cells), we sought to assess both direct effects and effects that are mediated by changes in cellular composition, as the genomic location and magnitude of mediated effects can inform us about how cell differentiation is regulated in response to environmental exposures. We thus measured, in all donors, the proportions of 16 immune cell subsets by standardized flow cytometry, including neutrophils, basophils, eosinophils, monocytes, natural killer (NK) cells, dendritic cells, B cells, CD4+ CD8- T cells and naïve, central memory (CM), effector memory (EM) and terminally differentiated effector memory cells (EMRA) CD4+ and CD8+ T cells.

We first determined which immune cell populations most affect DNA methylation variation, by quantifying differences in SmC levels between the 16 blood cell subsets with multivariable regression models including log-ratios of cell subsets, defined according to the hierarchical and compositional nature of the data (Methods). We verified that our models are accurate, using simulations and comparisons with independent DNA methylation data from sorted cellular subsets. We found that our estimated effects of cell subset log-ratios on SmC levels perform as expected on simulated data (Supplementary Fig. 2 and Supplementary Notes) and are highly correlated with DNA methylation differences observed between sorted immune cell fractions (R > 0.6; Supplementary Data 2). When applying these models on our data, we found that SmC levels of 134,079 CpG sites (20.8% of CpG sites, Supplementary Data 2) are associated with the log-ratio of myeloid vs. lymphoid lineages (Bonferroni corrected Padj < 0.05). Furthermore, the log-ratio of these subsets is the factor most associated with the first three Principal Components (PCs) of the DNA methylation data (multiple linear mixed model of PCI: P = 5.0 × 10^{-18}; PC2: P = 1.6 × 10^{-4}; PC3: P = 6.7 × 10^{-7}), which respectively explain 11.4, 7.5, and 5.5% of variation in DNA methylation. Importantly, we also found that 20,780 and 44,919 CpG sites are associated with the log-ratios of naïve and differentiated (CM, EM and EMRA) CD4+ and CD8+ T cell subsets, respectively (Padj < 0.05; Supplementary Data 2), supporting the view that SmC levels differ substantially among T cell subpopulations. Furthermore, the log-ratios of naïve and differentiated CD4+ and CD8+ subsets are also associated with PC1 and PC3 (P < 1.2 × 10^{-4}; Fig. 1c, d). These results indicate that differences in the proportion of naïve and differentiated subsets of CD4+ and CD8+ T cells contribute substantially to DNA methylation variation and may mediate associations between DNA methylation and environmental exposures or diseases.

#### Cell mixture deconvolution methods partially account for blood cell heterogeneity

Direct effects of environmental exposures or diseases on DNA methylation are often estimated by adjusting EWAS on major cell-type fractions, which are predicted in silico from SmC levels with cell mixture deconvolution methods. However, standard methods only predict the overall proportions of CD4+ and CD8+ T cells and may therefore overestimate the direct effects on DNA methylation of factors that affect T cell composition, such as aging and viral infections. To test this hypothesis, and to assess more generally how intrinsic and environmental factors affect the DNA methylome, we conducted EWAS of 141 candidate factors, by using linear mixed models adjusted on batch variables, genetic factors (i.e., associated meQTL variants), genetic ancestry, smoking status,
sex and a non-linear age term (Methods). Models were adjusted, or not, for the 16 measured cell proportions, to estimate total (i.e., direct and mediated) or direct effects, respectively. Mediated effects were estimated by mediation analysis (Methods). We considered that each EWAS constitutes a separate family of association tests and used the Bonferroni correction for multiple testing adjustment ($P_{\text{adj}} < 0.05$).

Out of the 141 candidate factors, those that have significant total effects on DNA methylation include age (n = 97,219 CpG sites; 15.1% of CpG sites), cytomegalovirus (CMV) serostatus (n = 79,654; 12.4%), sex (n = 23,002; 3.6%), heart rate (n = 2,924; 0.5%), smoking (n = 839; 0.1%), body temperature (n = 175), C-reactive protein (CRP) levels (n = 53), the hour of blood draw (n = 36) and traits related to lipid metabolism (n = 3; Fig. 1b and Supplementary Data 1). Accordingly, the first PCs of DNA methylation are most strongly associated with CMV (PC1: $P = 8.3 \times 10^{-13}$; PC2: $P = 7.8 \times 10^{-10}$), age (PC3: $P = 5.7 \times 10^{-29}$), and sex (PC4: $P = 2.2 \times 10^{-5}$), when not considering immune cell fractions (Fig. 1c, d and Supplementary Fig. 1i, j). When adjusting on blood cell composition, factors that have significant direct effects on DNA methylation include age (n = 35,701; 5.5%), sex (n = 17,067; 2.6%), smoking (n = 428; 0.07%), CMV serostatus (n = 243; 0.04%), CRP levels (n = 39) and lipid metabolism-related traits (n = 3; Fig. 1b, Supplementary Table 1)
Cytomegalovirus infection alters the blood DNA methylome through the regulation of host transcription factors

We identified CMV serostatus as one of the exposures that is associated with the largest number of CpG sites (Fig. 1b), CMV is the causative agent of a latent, mainly asymptomatic, infection that ranges in seroprevalence from 30 to 100% across populations. CMV is known to drastically alter the composition of the CD4+ and CD8+ T cell compartments in blood. Accordingly, we found that 85,922 CpG sites show a significant cell-composition-mediated effect of CMV serostatus on DNA methylation (adj P < 0.05; Supplementary Data 1), indicating that the effects of the latent infection are mainly mediated by cellular composition. Furthermore, we observed a strong correlation between measured and total effect sizes of CMV serostatus (R = 0.93; Fig. 1a) and 99.5% of CpG sites with a significant direct effect also show a significant mediated effect (n = 244/245). We found that mediated effect sizes of CMV are strongly correlated with estimated measures of DNA methylation differences between naïve and memory CD4+ and CD8+ T cells (R = 0.68 and R = 0.53, respectively; Fig. 1b), suggesting that cell-composition-mediated effects of CMV are predominantly attributable to changes in these T cell subsets.

One of the strongest cell-composition-mediated effects of CMV infection was observed in an intron of DNM3 A (β value scale 95% confidence interval [Cl]: [1.8%, 2.4%], Padj = 1.1 × 10−3), encoding a key DNA methyltransferase playing a role in the replication of some herpesviruses. CMV donors show a substantial increase in the proportion of CD4+ and CD8+ T cells (P = 6.8 × 10−5 and P = 1.9 × 10−5, respectively), which in turn are associated with higher 5mC levels at DNM3 A (P = 3.3 × 10−3 and P = 1 × 10−3, respectively), supporting mediation by differentiated T cell subsets (Fig. 2c). To test if the effects of CMV infection on 5mC levels are cell-type-dependent, we derived and verified an interaction model similar to CellDMC (Methods). We restricted this analysis to interactions with the proportion of cells from the myeloid lineage, as previously reported, and found only one CpG site where CMV effects depend on the proportion of myeloid cells (Padj < 0.05; Supplementary Data 3). These results indicate that CMV infection affects a large fraction of the blood DNA methylome primarily through changes in blood cell proportions, rather than through cell-type-dependent changes.

However, when adjusting for blood cell composition, including CD4+ and CD8+ T cell sub-types, a significant direct effect of CMV serostatus was detected for 245 CpG sites. Increased 5mC levels in CMV donors localize predominantly in enhancers and regions flanking transcription start sites (odds ratio [OR] > 3.0, Padj < 5.3 × 10−5; Supplementary Fig. 5), suggesting dysregulation of host gene expression as a result of latent infection. The second strongest direct effect of CMV infection was observed nearby the TS of LTBP3 (β value scale 95% CI: [1.9%, 3.1%], Padj = 7.1 × 10−3; Fig. 2d and Supplementary Fig. 6). LTBP3 is a regulator of transforming growth factor β (TGF-β), which is induced in CMV latently infected cells. Strikingly, CpG sites showing
increased 5mC levels in CMV+ donors are strongly enriched in binding sites for the BRD4 transcription factor (TF) (n = 187/189, OR = 48.0, 95% CI: [13.8, 42.2]; Fig. 2e and Supplementary Data 4), a bromodomain protein that plays a critical role in the regulation of latent and lytic phases of CMV infection\(^\text{44}\). Conversely, CpG sites showing a decrease in DNA methylation in CMV+ donors are strongly enriched in binding sites for BATF3 (OR = 1.07, 95% CI: [1.02, 1.15]; \(P_{\text{adj}} < 0.01\); Fig. 2f), which is paramount in the priming of CMV-specific CD8+ T cells by cross-presenting dendritic cells\(^\text{45}\). Collectively, these analyses imply that CMV infection directly affects the human blood DNA methylome through the dysregulation of host TFS implicated in viral latency and host immune response.

Finally, to motivate future research on the epigenetic effects of CMV infection, we used elastic net regression and stability selection to predict CMV serostatus from DNA methylation (Methods). Based on 547 CpG sites, the model predicts CMV serostatus with an out-of-sample accuracy of 87%, using 10-fold cross-validation. We anticipate that this model will be useful to determine if latent CMV infection can

**Fig. 2 | Effects of cytomegalovirus infection on the blood DNA methylome.**

**a** Total effects against cell composition-mediated effects of CMV infection on 5mC levels. **b** Effects of CD4+ T cell differentiation on 5mC levels against cell composition-mediated effects of CMV infection on 5mC levels. **c** Proportion of CD8+ T cells in CMV+ and CMV- donors (left panel). 5mC levels at the DNMT3A locus against the proportion of CD8+ T cells (right panel). 5mC levels are given in the \(\beta\) value scale. The black line indicates the linear regression line. **d** Genomic distribution of direct effects of CMV infection at the LTBP3 locus. **e** Enrichment of CpG sites with a significant direct, positive effect of CMV infection in binding sites for TFs. **f** Enrichment of CpG sites with a significant direct, negative effect of CMV infection in binding sites for TFs. **a** Only CpG sites with a significant cell composition-mediated effect are shown. The black line indicates the identity line. Tests were corrected for multiple testing by the Bonferroni adjustment. **a**, **b**, **d** Effect sizes are given in the M value scale. **e** The 15 most enriched TFs are shown, out of 1165 tested TFs. The point and error bars indicate the odds-ratio and 95% CI. CIs were estimated by the Fisher’s exact method. Statistics were computed based on a sample size of \(n = 884\) and for 644,517 CpG sites.

**Aging elicits DNA hypermethylation related to Polycomb repressive complexes and increased epigenetic dispersion**

Although the effects of aging on DNA methylation are well established\(^\text{35,36}\), it remains unclear the extent to which they are due to changes in unmeasured proportions of differentiated T cells (Fig. 1b) or CMV infection, which are both strongly associated with age\(^\text{25,32}\). Indeed, age has a significant total effect on 5mC levels at 97,219 and 113,742 CpG sites, when adjusting or not on CMV serostatus, and CMV infection mediates a substantial fraction of total age effects (\(n = 10,074\) CpG sites). We thus investigated how the blood DNA methylome is shaped by the intertwined processes of cellular aging (i.e., direct effects) and age-related changes in blood cellular composition (i.e., mediated effects), while accounting for CMV serostatus.

We found that, out of the 35,701 CpG sites associated directly with age, more than 97% were associated with age in a previous EWAS\(^\text{35}\), indicating a strong overlap (OR 95% CI: [35.6, 40.8]). In line with previous findings\(^\text{35,36}\), direct effects of age are typically larger than mediated effects (Fig. 3a). Furthermore, the strongest direct age effects, such as those observed at ELOVL2 and FHL2 (Supplementary Fig. 6), are not mediated by cellular composition (\(P_{\text{adj}} = 1.0\)), suggesting that age-related changes at these CpG sites are typically shared across cell-types. We observed that 61% of the CpG sites directly associated with age show a decrease in 5mC levels. Age-associated demethylation predominates outside of CpG islands (CGIs) and in regions flanking transcription start sites and in enhancers (Fig. 3b and Supplementary Fig. 7a, b). Conversely, DNA hypermethylation was observed in 95% of age-associated CpGs within CGIs. Consistently, CpG sites exhibiting increasing 5mC levels with age are mainly found in Polycomb-repressed regions, bivalent TSSs, and bivalent enhancers (Fig. 3b, c), which are CGI-rich.
Fig. 3 | Direct and cell-composition-mediated effects of aging on the blood DNA methylome. a Direct effects against cell-composition-mediated effects of age on 5mC levels (50-year effect). Only CpG sites with a significant direct or cell-composition-mediated effect are shown. Labels denote genes with strong direct or cell-composition-mediated effects of age. b Enrichment in CpG sites with significant direct effects of age, across 15 chromatin states. c Enrichment of CpG sites with a significant positive, direct effect of age in binding sites for TFs. d Increased variance of the proportion of CD4+ TEMRA cells with age. e Enrichment in CpG sites with significant direct effects of age, across 12 chromatin states. f Enrichment of CpG sites with significant cell-composition-mediated effects of age, across 12 chromatin states. g Enrichment of CpG sites with significant cell-composition-mediated effects of age, across 12 chromatin states. h Distributions of significant cell-composition-mediated effects of age, across 12 chromatin states. i Enrichment of CpG sites with significant cell-composition-mediated, positive effects of age in binding sites for TFs. j Numbers on the right indicate the number of associated CpG sites and proportion of positive effects. k The point and error bars indicate the odds-ratio and 95% CI. The 15 most enriched TFs are shown, out of 1165 tested TFs. l The 15 most enriched TFs are shown, out of 1165 tested TFs. m The 15 most enriched TFs are shown, out of 1165 tested TFs.
regions (Supplementary Fig. 1M, N). Furthermore, these CpG sites are most enriched in binding sites for RING1B, JARID2, RYBP, PCGF1, PCGF2, and SUZ12 TFs (OR > 10.0; Fig. 3d and Supplementary Data 4), which are all part of the Polycomb repressive complexes 1 and 2. PRC1 and PRC2 mediate cellular senescence and modulate longevity in invertebrates16-18. Importantly, when restricting the analysis to CpG sites outside of CpG islands, we found similar enrichments in Polycomb-repressed regions (OR 95% CI [17.7, 20.0]) and PRC TF binding sites (RING1B OR 95% CI [19.9, 22.4]; PCGF2 OR 95% CI [17.8, 20.7]). Finally, genes with age-increasing Smc levels are strongly enriched in developmental genes (Padj = 1.7 × 10^{-46}; Supplementary Data 5), which are regulated by PRCs17. Overall, these results confirm previously described effects of age on the blood DNA methylome, while accounting more comprehensively for blood cell composition and CMV infection, and support a key regulatory role of Polycomb proteins in age-related hypermethylation18.

We then assessed whether age-related changes in blood cell composition or CMV seropositivity could contribute to age-related changes in the variance of Smc levels, a phenomenon known as “epigenetic drift” (i.e., the divergence of the DNA methylome as a function of age owing to stochastic changes)19. We observed that the proportion of several cell types in the blood is increasingly dispersed with aging, such as CD4+ TEMRA cells (Fig. 3e). Therefore, we then assessed whether age-related changes in blood cell composition or CMV seropositivity could contribute to age-related changes in the variance of Smc levels, a phenomenon known as “epigenetic drift” (i.e., the divergence of the DNA methylome as a function of age owing to stochastic changes)19. We observed that the proportion of several cell types in the blood is increasingly dispersed with aging, such as CD4+ TEMRA cells (Fig. 3e). Therefore, we observed a significant dispersion of DNA methylome with age for 3.1% of all CpG sites (Padj = 0.05). We compared these CpG sites with those previously reported to be increasingly variable with age in whole blood and monocytes20 and replicated 2604 of 5,075 CpG sites, supporting a strong overlap between the two different approaches (OR 95% CI [36.2, 40.5]). An example of a CpG site with a large, age-increasing dispersion is found in the TSS of MAF (Padj = 4.4 × 10^{-43}; Fig. 3f), encoding a transcription factor that regulates insulin. Strikingly, 99.4% of CpGs with age-related dispersion show an increase in the variance of Smc levels with age (Fig. 3g), supporting a decrease in the fidelity of epigenetic maintenance associated with aging. In addition, we found that, out of 20,140 CpG sites with age-related dispersion, 87.3% show no significant changes in mean Smc levels with age, and we detected no correlation between estimates of dispersion and direct age effect sizes (Fig. 3h), implying that these results are not driven by relationships between the average and variance of Smc levels. Furthermore, when also adjusting the variance function for cellular composition, we found evidence of dispersion in 8,576 CpG sites (Padj < 0.05), with similar effect sizes as in the previous model (R = 0.93; Methods). 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 -1.1% of CpG sites (n = 7090; Fig. 3a and Supplementary Data 1), indicating that a substantial fraction of age-related changes in DNA methylation are due to age-related changes in immune cell proportions. Mediated effects are most often associated with demethylation (76% of age-associated CpG sites), regardless of the chromatin state or CGI density of the loci considered (Fig. 3 and Supplementary Fig. 7c, d). Enhancers and regions flanking transcription start sites are enriched in CpG sites with a significant cell-composition-mediated effect of age (Fig. 3i), possibly because these regions tend to be regulated in a cell-type-dependent manner21. In contrast with direct age effects, CpG sites with a cell-composition-mediated increase in DNA methylation are enriched in TF binding sites for RUNX1-3 (OR = 8.5; 95% CI: [4.5, 14.7]; Padj < 1.2 × 10^{-4}), which are key regulators of hematopoiesis (Fig. 2k and Supplementary Data 4). Genes with CpG sites showing a mediated increase or decrease in DNA methylation with age are enriched in genes involved in lymphoid (Padj = 2.0 × 10^{-7}), myeloid (Padj = 6.1 × 10^{-7}) cell activation, respectively (Supplementary Data 5). This indicates that mediated effects of age on DNA methylation are related to progressive, lifelong differences in the composition of the lymphoid and myeloid cell lineages.

We then determined if age effects on Smc levels depend on the proportion of cells from the myeloid lineage, by using an interaction model (Methods). In line with a previous study22, we found that cell-type-dependent effects of age (Supplementary Data 3) are limited; only 10 CpG sites show DNA methylation changes with age that depend on the proportion of myeloid cells (Padj < 0.05; Supplementary Data 3). Importantly, age also has a strong mediated effect on all these CpG sites (Padj < 1.0 × 10^{-20}), implying that these loci are associated with age because of changes in blood cell composition, although their relation to age is cell-type-dependent. Collectively, our findings provide substantial evidence that DNA methylation variation with age results from different, non-mutually exclusive mechanisms: the progressive decline of the epigenetic maintenance system that is common to all cell types, the increased heterogeneity of immune cell subsets that characterizes immunosenescence23 and, to a lesser extent, accelerated changes within specific blood cell compartments.

**Sex differences in DNA methylation are predominantly cell- and age-independent**

Given that substantial differences in immune cell composition have been observed between women and men24, we next assessed how cellular heterogeneity contributes to sex differences in DNA methylation10-13. We found 3.6% of CpG sites (n = 23,002) with a significant total effect of sex, 2.6% (n = 17,067) with a significant direct effect, and only 0.2% (n = 1385) with a significant cell-composition-mediated effect (Padj < 0.05; Supplementary Fig. 5a and Supplementary Data 1). Out of CpG sites directly associated with sex, 96.2% were already associated with sex in a previous EWAS13, indicating again a strong overlap (OR 95% CI: [39.6, 46.5]). The largest direct effects of sex were observed at Dyrk2, Dnmi, Rftn1, Hydin, and Nabi genes (Padj < 1.0 × 10^{-20}; Supplementary Fig. 6). For example, the Dyrk2 promoter is 11.7% and 45.6% methylated in men and women, respectively, at a CpG site that we found to be bound by the X-linked PHF8 histone demethylase (Supplementary Fig. 8b, c). Dyrk2 phosphorylates amino acids and plays a key role in breast and ovarian cancer development25.

DNA methylation levels are higher in women at 79.7% of sex-associated autosomal CpG sites (Supplementary Fig. 8d, e), a pattern also observed in newborns26. This proportion is similar across different genomic regions, based on either chromatin states or CpG density (Supplementary Fig. 8e, g). When quantifying how sex differences in DNA methylation vary during adulthood, by adding a sex-by-age interaction term to our models (Methods), we found only 7 CpG sites with a significant, sex-dependent effect of age (Padj < 0.05; Supplementary Data 3). Confirming previous findings13,16, the strongest sex-by-age interaction effects were found at Fign (Padj = 7.1 × 10^{-7}), associated with risk-taking behaviors27 and educational attainment26, and Prr4 (Padj = 5.6 × 10^{-7}), associated with the dry eye syndrome, a hormone-dependent, late-onset disorder28. 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.
Gene × cell type and gene × environment interactions affect DNA methylation variation

Gene × environment interactions are thought to underlie adaptable human responses to environmental exposures through epigenetic changes. To test if gene × environment interactions affect DNA methylation, we first estimated, for each CpG site, the effects on 5mC levels of local and remote DNA sequence variation, defined as genetic variants within a 100-Kb window and outside a 1-Mb window, centered on the CpG site, respectively (Methods). We considered local and remote meQTLs to be independent families of tests and used the Bonferroni correction to adjust for multiple testing. We found a significant local meQTL for 107,048 CpG sites and a significant remote meQTL for 1228 CpG sites \((P_{\text{adj}} < 0.05\); Supplementary Fig. 9 and Supplementary Data 6). In agreement with previous studies, CpG sites with a local meQTL are enriched in enhancers (OR 95% CI: [2.09, 2.21]) and depleted in TSS and actively transcribed genes (OR 95% CIs: [0.52, 0.56] and [0.57, 0.60]; Fig. 4a). Conversely, CpG sites under remote genetic control are enriched in TSS regions (OR 95% CI: [2.10, 2.21]) and depleted in TSS and actively transcribed genes (OR 95% CIs: [0.05, 0.07], [0.60, 0.67]; Fig. 4b).

We next explored whether effects of genetic variants on 5mC levels depend on the circulating proportion of myeloid cells. We found evidence for cell-type-dependent meQTLs at only 249 CpG sites \((P_{\text{adj}} < 0.05\); Fig. 4d and Supplementary Data 3), supporting the notion that genetic effects on 5mC levels are generally shared across blood cell subsets. The strongest signal was found between 5mC levels upstream of CLEC4C and the nearby rs11055602 variant, which has been previously shown to strongly affect CLEC4C protein levels. This C-type lectin, known as CD303, is used as a differentiation marker for dendritic cells, suggesting the epigenetic regulation of the locus is cell-type-dependent. Accordingly, rs11055602 genotype effects on DNA methylation depend on the circulating proportions of myeloid cells (beta scale interaction effect, 95% CI: [0.16, 0.22], \(P_{\text{adj}} = 7.4 \times 10^{-7}\); Fig. 4e), and dendritic cells (95% CI: [-8.3, -5.0], \(P_{\text{adj}} = 3.5 \times 10^{-13}\)).

We then evaluated whether the main non-heritable determinants of DNA methylation variation in our cohort, i.e., age, sex, CMV serostatus, smoking status and chronic low-grade inflammation (CRP levels; Fig. 1b, Supplementary Fig. 3 and Supplementary Notes), can affect 5mC levels in a genotype-dependent manner. We thus tested for genotype × age, genotype × sex, genotype × smoking jointly (Methods). Genotype × CRP levels interactions were tested in separate models that also include the other interaction terms. We found statistical evidence for genotype-dependent effects of age and sex at 68 and 20 CpG sites, respectively \((P_{\text{adj}} < 0.05\); MAF > 0.10; Fig. 4d and Supplementary Data 3), the interacting meQTL variant being local in all cases. We detected a strong genotype × age interaction for two 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 (beta scale 95% CI: [0.11, 0.13], \(P_{\text{adj}} = 7.28 \times 10^{-10}\); Fig. 4f). 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. Another strong genotype × age interaction effect was found for a CpG site upstream of FCERIA, encoding the high-affinity IgE receptor. FCERIA 5mC levels decrease with age in rs2231746 T > C carriers only (95% CI: [0.05, 0.07], \(P_{\text{adj}} = 8.6 \times 10^{-12}\)), a variant known to control serum IgE levels. Collectively, our analyses identify few, albeit strong, environment- and cell-type-dependent meQTLs, supporting the relatively limited impact of gene × cell type and gene × environment interactions on the blood DNA methylation.

Cellular composition and genetics drive DNA methylation variation in human blood

Having established how cellular composition, intrinsic factors, genetic variation, and a broad selection of non-heritable factors shape the blood DNA methylene, we next sought to compare the relative impact of these factors on DNA methylation. We classified the factors into four groups: (i) the cellular composition 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 CRP levels. 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 (Methods). To ensure unbiased estimates, we mapped local meQTLs anew within each training set.

The full model that includes all groups explains ~5% of out-of-sample variance for 52.3% of CpG sites (Fig. 5a), which are typically characterized by low total 5mC variance (Supplementary Fig. 10). This suggests that these sites are constrained in the healthy 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 20.8% of CpG sites \((n = 134,305)\). The strongest predictor for these CpG sites is cellular composition, genetics, intrinsic factors and exposures in 74.7%, 21.3%, 3.8% and 0.01% of cases, respectively. Cellular composition explains ~25% of out-of-sample variance for 1.0% of CpG sites \((n = 90,033);\) Fig. 5a, c and Supplementary Data 7), with the highest variance explained by cellular composition for one CpG site being 71.8%. For the 2,580 CpG sites where the model explains >75% of variance, local DNA sequence variation is the strongest predictor in 99.2% of cases (Fig. 5c and Supplementary Data 7). Local genetic variation explains ~25% of DNA methylation variance at 23,677 CpG sites, and almost as many when adjusting for cellular composition \((n = 22,865);\) (Fig. 5a, b), indicating that genetic effects on 5mC levels are mainly cell-type-composition-independent. Intrinsic factors explain ~25% of out-of-sample variance at 3669 CpG sites, and >75% at 16 sites (Fig. 5c). When conditioning on cell composition, these numbers dropped to 334 and 6 CpG sites, respectively, suggesting that the predictive ability of age and sex is partly mediated by immune cell composition (Fig. 5b).

Interestingly, environmental exposures are the weakest predictor of 5mC levels, explaining ~25% of the variance at only 29 CpG sites and with a maximum variance explained for a CpG site of 50.1%.

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 (Methods). We found a significant increase in predictive ability when including interaction terms for 431 CpG sites (ANOVA \(P_{\text{adj}} < 0.05\)). However, the effects were typically modest: only 13 CpG sites showed an increase in the proportion of variance explained larger than 5% (Fig. 5b). 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 blood cellular composition, age, sex, genetics, environmental exposures, and their interactions to variation in the DNA methylene. All the results can be explored via a web-based browser (http://mimeth.pasteur.fr/), to facilitate the exploration of the estimated effects of these factors on DNA methylation variation. We found that CMV infection elicits substantial changes in the blood DNA methylation, in line with their role in the regulation of amyloid beta peptide, a critical component in the etiology of Alzheimer’s disease.

Nature Communications | (2022) 13:5895 - 8
methyloeme, in contrast with other herpesviruses such as EBV, HSV-1, HSV-2, and VZV. Latent CMV infection is known to profoundly alter the number, activation status and transcriptional profiles of immune cell populations, yet its epigenetic consequences have attracted little attention. We observed that most CMV effects on DNA methylation are mediated by the profound changes in blood cell composition, including the CMV-driven inflation of memory CD4+ and CD8+ T cells. However, we also detected cell-composition-independent effects of CMV infection, suggesting that the herpesvirus can directly regulate the host epigenome. Notably, differentially methylated CpG sites in CMV+ donors are strongly enriched in binding sites for BRD4, a key host regulator of CMV latency, suggesting that the recruitment of BRD4 by CMV during latent infection affects BRD4-regulated host genes. Furthermore, CMV+ donors are characterized by a strong increase in 5mC levels at LTBP3, the product of which is involved in TGFβ secretion. TGFβ is a well-known immunosuppressive cytokine induced by CMV infection, 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.

Our study provides further support to the notion that three different biological mechanisms contribute to age-related changes in DNA methylation. The first elicits an increased dispersion of 5mC levels with age that is related to epigenetic drift. We found that dispersion of DNA methylation with age is not due to cellular heterogeneity, supporting instead the progressive decline in fidelity of the DNA methylation maintenance machinery across cell populations. The second mechanism results in cell-composition-independent, global DNA demethylation and CGI-associated hypermethylation. Age-associated DNA demethylation could be related to the down-regulation 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.0.6-0.8. 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.0.9-0.11. Both scenarios are supported by the enrichment of Polycomb-repressed regions in age-associated CpG sites, and 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 genome, particularly at enhancers of myeloid activation genes. This process likely reflects an increased degree of differentiation in the lymphoid compartment with age. Single-cell methylation of differentiating and dividing white blood cells will help determine the role of mitotic and post-mitotic 5mC changes during epigenetic aging.

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. Twin studies have estimated the heritability of DNA methylation to range from ~20–40% (ref. 86-88), suggesting that environmental effects, along with gene × environment interactions, account for the remaining 60–80% (ref. 89). 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% (ref. 90). Nevertheless, a limitation of our study is that perinatal and early life exposures, which are thought to contribute extensively to epigenetic variation in adulthood, 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, but statistical evidence for interaction effects requires larger cohorts, suggesting that our results might represent a small, perceptible fraction of a large number of weak effects.0.12-0.14. Large, longitudinal 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, our analyses show that first-generation cell mixture deconvolution
methods do not fully distinguish direct from cell-composition-mediated effects of CMV infection and age on DNA methylation, probably because these two factors alter the proportions of blood cell subsets that are not estimated by these methods. This reinforces the view that EWAS must be interpreted with great caution, particularly when the studied diseases or conditions are known to affect unmeasured immune cell fractions. Encouragingly, our findings suggest that, when blood cell composition is not measured directly, high-resolution cell mixture deconvolution methods provide a more complete correction for cellular heterogeneity and are therefore expected to improve the interpretation of future epigenomic studies. Second, because age, sex, CMV infection, smoking, and chronic low-grade inflammation can influence disease risk, our results emphasize the critical need to consider such factors in EWAS, as these factors can confound associations. Lastly, our findings reveal the epigenetic impact of aging and persistent viral infection through fine-grained changes in blood cell proportions, highlighting the need to assess the respective role of altered cellular composition and DNA methylation in the etiology of disease. 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.

Methods

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 1000 healthy donors of the Milieu Intérieur cohort were recruited by BioTrial (Rennes, France), and included 500 women and 500 men. All subjects provided written informed consent, including for genetic studies, prior to enrollment in the study. 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. 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 11 April 2018.

DNA sampling and extraction

Whole blood was drawn from the 1000 Milieu Intérieur healthy, fasting donors every working day from 8AM to 11AM, from September 2012 to August 2013, in Rennes, France. Different anticoagulants were used, depending on the downstream analyses. For DNA methylation profiling, blood samples were collected on EDTA, whereas samples for flow cytometry and genome-wide DNA genotyping were collected on Litheparin. Tracking procedures were established in order to ensure delivery to Institut Pasteur (Paris) within 6 h of blood draw, at a temperature between 18 °C and 25 °C. Upon receipt, samples were kept at room temperature until DNA extraction. DNA for DNA methylation profiling was extracted using the Nucleon BACC3 genomic DNA extraction kit (catalog #: RPN8512; Cytiva, Massachusetts, USA). High-quality genomic DNA was obtained for 978 out of the 1000 donors.

DNA methylation profiling and data quality controls

Extracted genomic DNA was treated with the EZ DNA Methylation Kit (catalog #: D5001; Zymo Research, California, USA). Bisulfite-converted DNA was applied to the Infinium MethylationEPIC BeadChip (catalog #: WG-317–1003; 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. All samples

![Figure 5](https://example.com/figure5.png)

**Fig. 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 (i.e., age and sex) and environmental exposures (i.e., smoking, CMV infection and CRP levels), for 644,517 CpG sites. b Complementary cumulative distribution function of the out-of-sample variance explained by genetic factors, intrinsic factors, environmental 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 environmental exposures, respectively.
showed average detection $P$-values $<$ 0.005. No sample showed a mean of methylated intensity signals lower than $3 \times$ standard deviations (SD) from the cohort average. Therefore, 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 (Supplementary Fig. 1a). Using the 59 control SNPs included in the MethylationEPIC array, a single sample showed high genotypic discordance with the genome-wide SNP array data (see ‘Genome-wide DNA genotyping’ section) and was thus excluded (Supplementary Fig. 1b). Unmethylated and methylated intensity signals were converted to $M$-values. A total of 2930 probes with $>1\%$ missingness (i.e., $P$-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 irba R package, Principal Component Analysis (PCA) of $M$ values identified nine outlier samples, including eight that were processed on the same array (Supplementary Fig. 1c), which were also excluded. The ‘noob’ background subtraction method$^{(1)}$ was applied on $M$ values for the remaining 968 samples, which showed highly consistent epigenome-wide DNA methylation profiles (Supplementary Fig. 1d, 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 age, sex and cytomegalovirus (CMV) serostatus as fixed effects, and slide position and sample plate as random effects. The models were fitted with the lmme4 R package$^{(2)}$. Strong associations were observed between the first four PCs and slide position and sample plate (Supplementary Fig. 1f, g).

$M$-values were converted to $\beta$ values, considering that $\beta = 2^{M/2} \left(2^{-M} + 1\right)$.

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 $\beta$ values that were greater than $5 \times$ SD from the population average, corresponding to $<$0.1% of all measures. We also excluded (i) 83,350 non-specific probes that share $>90\%$ sequence identity with several genomic regions (see details in$^{(3)}$), (ii) 118,575 probes that overlap a SNP that is within the 50 pb surrounding the CpG site and has a MAF $>1\%$ in the Milieu Intérieur cohort or in European populations from the 1000 Genomes project$^{(4)}$, (iii) 558 probes that were absent from the Illumina annotations version 1.0 B4 and (iv) 16,576 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**

Immune cell proportions were measured using ten eight-color flow cytometry panels$^{(5)}$. 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 counts were exported from FlowJo, Protocols, panels, staining antibodies, and quality control filters used for flow cytometry analyses are detailed elsewhere$^{(5)}$. Abnormal lysis or staining were systematically flagged by trained experimenters. We removed outliers by using a scheme detailed previously$^{(5)}$. 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).

Problems in flow cytometry processing, such as abnormal lysis or staining, were systematically flagged by trained experimenters, which resulted in 8.7% missing data. Because imputing missing data for donors who show large missingness could be inaccurate, we excluded 74 donors with no data for the T cell panel. Finally, the remaining missing data were imputed using the random forest-based missForest R package$^{(6)}$.

**Genome-wide DNA genotyping**

The 1000 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 combined data set included 732,341 high-quality polymorphic SNPs. After genotype imputation and quality-control filters$^{(7)}$, a total of 11,395,534 SNPs was further filtered for minor allele frequencies $>$5%, yielding a data set composed of 1000 donors and 5,699,237 SNPs for meQTL mapping. Ten pairs of first to third-degree related donors were detected with KING 1.9 (ref. $^{107}$). Out of the 894 donors whose blood methylome and blood cell composition were accurately profiled, 884 unrelated donors were kept for subsequent analyses.

**Immune cell proportions**

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 cell types or if they in fact reflect changes in blood cell composition. To answer this question, we considered the 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$^{(8)}$. As these cellular proportions were measured by flow cytometry using a hierarchical gating strategy$^{(9)}$, they are expected to sum to one. Yet, because of measurement errors, cell fractions do not exactly sum to one in all donors. For a measure of the proportion of a given cell subset in a given donor, we therefore used the absolute count of the cell type divided by the sum of absolute counts of all the 16 measured cell subsets. We used the same approach when considering a reduced set of six major cell types, comprising neutrophils, monocytes, NK cells, B cells, and CD4+ and CD8+ T cells, for comparison purposes.

**Compositional analysis of cellular component**

We sought to study the association between 5mC levels and blood cell composition, experimentally measured by flow cytometry. However, the 16 measured cellular proportions are constrained to be positive and to sum to one. Consequently, a change in one cellular proportion must necessarily change one or more of the other cellular proportions, complicating the interpretation of parameters estimated from linear regression models with measured immune cell proportions as predictors$^{(10,11)}$. Here, we investigated instead the effect of balances, which are transformations of cell-type proportions that can be seen as a generalization of the logit-transform. These balances model the effect of a relative change between two groups of cell types. They are defined in a hierarchical manner of increasing granularity, by a sequential binary partition (SBP) of the 16 measured cell types, generating 15 balances in total (Supplementary Data 2). As an example, we describe the first two balances. The other balances are defined in an analogous manner according to the SBP and the general procedure detailed elsewhere$^{(12)}$. The first balance captures the relative effect on
SmC levels of the myeloid cell types compared to the lymphoid cell types. Of the 16 measured cell types, five are myeloid and eleven are lymphoid. Let \( c^{\text{ML}}_i \ldots c^{\text{L}}_i \) be the measured myeloid proportions and \( \epsilon^{\text{ML}}_i \ldots \epsilon^{\text{L}}_i \) be lymphoid proportions for the \( i \)th individual. The first balance predictor for that individual is defined by

\[
b^i_1 = \frac{5}{5+\Pi} \log \left\{ \frac{\sum_{k=1}^{15} c^{\text{ML}}_k}{\sum_{k=1}^{15} c^{\text{L}}_k} \right\}.
\]

The second balance is defined within the lymphoid group and captures the relative effect on SmC levels of T cells with respect to NK observed value of the section performed an epigenome-wide association analysis of cellular composition on SmC levels at individual CpG sites (see section section‘Epigenome-wide association study of cell composition’). The 15 balances were used to investigate the effects of immune cell composition on SmC levels at individual CpG sites. For the \( i \)th individual on the \( k \)th CpG site, we used to estimate the effects of cellular composition on PCs of DNA methylation was,

\[
y^k_i = \mu^k + b^k P^{SNP} + \text{DateOfSampling}^k + \varepsilon^k_i,
\]

where \( \varepsilon^k_i \sim (0, \sigma^2_k) \). Models were fitted by ordinary least squares. For each balance in \( b_i \) (see Eqs. (1) and (2) for examples), the parameters in \( P^{SNP} \) are interpreted as the change in SmC levels for an increase in the first cell-type group and the corresponding decrease in the second cell-type group.

\[
y^k_i = \mu^k + b^k P^{SNP} + \text{DateOfSampling}^k + \varepsilon^k_i,
\]

\[
y^k_i = \mu^k + b^k P^{SNP} + \text{DateOfSampling}^k + \varepsilon^k_i,
\]

\[
y^k_i = \mu^k + b^k P^{SNP} + \text{DateOfSampling}^k + \varepsilon^k_i.
\]

Epigenome-wide association study of cell composition

To investigate how immune cell composition affects the blood DNA methylation, we investigated effects of cell-type balances on SmC levels at each CpG site. For the \( p \)th CpG site and the \( i \)th individual, introduce observed SmC levels \( y^p_i \) measured on the M value scale. Let \( b_i \) be a vector of 15 cell-type balances with corresponding parameter vector \( P^{SNP} \). Let the vector \( SNP^p \) contain the significant local SNP with the smallest \( P \)-value and all independently associated remote SNPs (see section ‘Local meQTL mapping analyses’ and section ‘Remote meQTL mapping analyses’) with corresponding parameter vector \( P^{SNP} \). We performed an epigenome-wide association analysis of cellular composition by fitting the models,

\[
y^p_i = \mu^p + b^p P^{SNP} + \text{DateOfSampling}^p + \varepsilon^p_i,
\]

where \( \varepsilon^p_i \sim (0, \sigma^2_p) \). Models were fitted by ordinary least squares. For each balance in \( b_i \) (see Eqs. (1) and (2) for examples), the parameters in \( P^{SNP} \) are interpreted as the change in SmC levels for an increase in the first cell-type group and the corresponding decrease in the second cell-type group.

Associations with principal components of DNA methylation

To evaluate how principal components (PCs) of DNA methylation levels are related to cell composition, we first computed PCs of SmC levels at 644,517 CpG sites, with the iriba R package. Let \( y^p_i \) be the observed value of the \( k \)th PC of the DNA methylation data and \( b_i \) a vector of 15 cell-type balances measured for individual \( i \) with the corresponding parameter vector \( P^{SNP} \). Given that we observed variability in SmC levels across dates of blood draw, we included them as random effects. Let \( \text{DateOfSampling}^p \) be the day of blood draw for the \( i \)th individual. The model we used to estimate the effects of cellular composition on PCs of DNA methylation was,

\[
y^p_i = \mu^p + b^p P^{SNP} + \text{DateOfSampling}^p + \varepsilon^p_i,
\]

where \( \varepsilon^p_i \sim (0, \sigma^2_p) \). Models were fitted by the lme4 R package. To evaluate how PCs of DNA methylation levels are related to the candidate non-heritable factors, i.e., age, sex, smoking status, CMV serostatus, introduce the variables Age, Woman, Exsmoker, Smoker, and CMV/ with corresponding parameters \( \beta^p_{\text{Age}} \), \( \beta^p_{\text{Woman}} \), \( \beta^p_{\text{Exsmoker}} \), \( \beta^p_{\text{Smoker}} \), and \( \beta^p_{\text{CMV}} \). Let PC1 and PC2 be the first two PCs of the genotype matrix. Let \( x_i \) be a vector of 15 measured cell proportions, excluding neutrophils because of the sum-to-one constraint, and \( P^{SNP} \) the corresponding parameter vector. The model we used to estimate the effects of non-genetic factors on PCs of DNA methylation was,

\[
y^p_i = \mu^p + c^p x_i + \text{Age} + \text{Woman} + \text{Exsmoker} + \text{Smoker} + \text{CMV} + \text{PC1} + \text{PC2} + \varepsilon^p_i.
\]

The models were fitted with the lme4 R package. Inference was performed using the Kenward-Roger F-test approximation for linear mixed models, implemented in the pbkrtest R package.

Epigenome-wide association studies of non-genetic factors

We assessed the effects of 141 non-genetic variables (Supplementary Data 1) on the blood DNA methylation of adults. The measured SmC levels at a CpG site are the average of the DNA methylation state at this CpG site of all cells in the blood sample. Many of the 141 candidate variables might influence cell composition, which will cause a corresponding change in SmC levels. We denote this effect the “cell-composition-mediated effect”. In addition, the variable might alter SmC levels within individual cells, or within cell-types. We denote this effect the direct effect (see Supplementary Fig. II for a schematic directed acyclic graph of the system). Several 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. 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 SmC levels induced by changes in cellular composition (i.e., cell-composition-mediated effects) and those induced within cell types (i.e., direct effects). For each variable of interest \( x \) and each CpG site, the total effect was estimated in a regression model including, as response variable, the SmC levels of the CpG site on the M value scale and, as predictors, \( x_i \), a nonlinear age term of 3 DoF natural splines, sex, CMV serostatus, smoking status, the significant local SNP with the smallest \( P \)-value, independently associated remote SNPs and the first two PCs of the genotype matrix. Again, since we observed variability in SmC levels across dates of blood draw, we included them as a random effect term. For the \( p \)th CpG site, let \( y^p_i \) be the SmC levels of the \( i \)th individual on the M value scale, \( f^p_{\text{Age}}(\text{Age}_i) \) a nonlinear age term of 3 DoF natural splines and \( SNP^p \) a vector of the minor allele counts for the significant local SNP with the smallest \( P \)-value and independently associated remote SNPs, with corresponding parameter vector \( P^{SNP} \). The total effect of the variable \( x_i \) was estimated by the corresponding parameter
\[ y_i^t = \beta_p^0 + x_i^t \beta_p + f_{\text{age}}^p(Age) + \text{Woman} \beta_p^\text{Woman} + \text{Exsmoker} \beta_p^\text{Exsmoker} + \text{Smoker} \beta_p^\text{Smoker} + \text{CMV} \beta_p^\text{CMV} + \text{PC1} \beta_p^\text{PC1} + \text{PC2} \beta_p^\text{PC2} + (\text{SNP}_j^p \beta_p^\text{SNP} + \text{DateOfSampling}_j^p + \epsilon^p), \]

where \( \text{DateOfSampling}_j^p - \mathcal{N}(0, \sigma^2_j^p) \) and \( \epsilon^p - \mathcal{N}(0, \sigma^2_p) \). The effect of aging was tested in models with \( x \) replaced by a linear one. The effects of sex, smoking status and CMV serostatus were tested in models where we removed \( x \). For variables relating to women only (e.g., age of menarche), we excluded men from the analysis and removed \( \text{Woman} \beta_p^\text{Woman} \). The models were fitted with the \texttt{Ime4 R package}^{10}. Hypothesis tests were performed using the Kenward-Roger approximation of the \texttt{F}-test for linear mixed models, implemented in the \texttt{pbkrtest R package}^{10}.

The direct effect. Let the vector \( c_i \) be measured proportions of the 15 immune cell types, excluding neutrophils, for the \( i \)th individual and \( \beta_p^c \) the corresponding parameter vector. Using the same notation as for the total effect, the direct effect of the variable \( x_i \) was estimated by \( \beta_p^c \) in the models,

\[ y_i^t = \mu^p + x_i^t \beta_p^c + c_i^t \beta_p^c + f_{\text{age}}^p(Age) + \text{Woman} \beta_p^\text{Woman} + \text{Exsmoker} \beta_p^\text{Exsmoker} + \text{Smoker} \beta_p^\text{Smoker} + \text{CMV} \beta_p^\text{CMV} + \text{PC1} \beta_p^\text{PC1} + \text{PC2} \beta_p^\text{PC2} + (\text{SNP}_j^p \beta_p^\text{SNP} + \text{DateOfSampling}_j^p + \epsilon^p), \]

We also tested the interaction effect of sex, CMV serostatus and smoking status with age by including one interaction term at a time in the model specified in Eq. (7). The models were fitted with the \texttt{Ime4 R package}^{10}. Hypothesis tests were performed by the Kenward-Roger approximation of the \texttt{F}-test for linear mixed models, implemented in the \texttt{pbkrtest R package}^{10}.

The mediated effect. The cell-composition-mediated effect was estimated as the effect on 5mC levels mediated by changes in proportions of the 16 cell subsets due to the given factor. We estimated the mediated effect of aging, sex, variables related to smoking, CMV serostatus and heart rate. The mediated effect was estimated using a two-stage procedure. First, we fitted models with measured proportions of immune cells as response variables. Let \( c_i \) be a vector of measured proportions of the 15 blood subsets, excluding neutrophils. Let \( c_i^p \) denote the \( p \)th entry of the vector \( c_i \), i.e., the measured proportion of the \( p \)th cell-type for the \( i \)th individual. Introduce the vector \( k_i \) of covariate values for the \( i \)th individual, including age (3 DoF spline with an entry for each term), sex, smoking, CMV serostatus and ancestry (2 PCs), but excluding the variable of interest \( x_i \) (mediated effect of aging was estimated with a linear term). For the model of the \( p \)th cell-type, let \( \beta_p^k \) be the parameter vector for the covariate vector \( k_i \) and \( \beta_p^c \) the parameter for the variable of interest \( x_i \). In the first stage, we fitted the models,

\[ E(c_i^t | x, k_i, \theta_p^c) = \beta_0 + x_i \beta_p^c + k_i \beta_p^k, \quad n = 1 \ldots 15. \]

Next, let \( y_i^t \) be 5mC levels in the M value scale for the \( p \)th CpG site, \( \theta_p^c \) the parameter for the variable of interest, and \( \theta_p^c \) and \( \theta_p^k \) parameter vectors for the effects of cell proportions and covariates. In the second stage, we fitted the models,

\[ E(y_i^t | x, k_i, \theta_p^c, \theta_p^k) = \beta_0 + x_i \theta_p^c + c_i^t \theta_p^c + k_i \theta_p^k. \]

The mediated effect of \( x_i \) on DNA methylation was estimated by \( \beta_p^c \theta_p^c \) (ref. 34). Inference was performed by the parametric bootstrap.

The direct effects adjusted by deconvolution methods. To compute the IDOL and Houseman-adjusted effects, we estimated proportions of CD4+ and CD8+ T cells, B cells, NK cells, monocytes, and neutrophils, using the estimateCellCounts2 function in the \texttt{FlowSorted.Blood.EPIC package} with either Houseman et al.’s CpG sites, or IDOL optimized CpG sites^{112}. For age, sex, smoking status, CMV serostatus, heart rate, ear temperature and hour of blood draw, we estimated the IDOL- and Houseman-adjusted effect by adjusting for estimated 5 proportions in the model specified by Eq. (7), instead of the 15 measured proportions, excluding neutrophils because of the sum-to-one constraint. To compute the EPIC IDOL-Ext-adjusted effects, we estimated proportions of 12 major cell types in blood, including CD4+ and CD8+ T cells, naive and differentiated subtypes of CD4+ and CD8+ T cells, neutrophils, monocytes, basophils, eosinophils, NK cells, regulatory T cells, naive and memory B cells, using the IDOL-Ext reference matrix in the \texttt{estimateCellCounts2} function from the \texttt{FlowSorted.BloodExtended.EPIC R package}^{112}. We estimated the IDOL-Ext-adjusted effect by including 11 estimated proportions in Eq. (7) instead of the 15 measured proportions, excluding neutrophils because of the sum-to-one constraint. Finally, for comparison purposes, we also computed the association between non-genetic factors and 5mC levels by adjusting, in Eq. (7), for the proportions of the 5 major cell types measured by flow cytometry, instead of the 15 measured proportions, excluding again neutrophils.

Prediction of CMV serostatus

We built a prediction model to estimate CMV serostatus from DNA methylation data using elastic net regression for binary data^{113}, implemented in the \texttt{glmnet R package}^{114}. We included all CpG sites as predictors in the model, including those on the X and Y chromosomes. The model was built from 863,906 CpG sites in 969 samples. The elastic net model has two tuning parameters that determine the degree of regularization of the predictor function. We selected both tuning parameters by two-dimensional five times repeated cross-validation over the two parameters. The final model fitted on the full data set includes 547 CpG sites with non-zero parameters.

Detection of the dispersion of DNA methylation with age

To estimate changes in dispersion of 5mC levels with age, we fitted regression models where the residual variance depends on age. Let \( y_i^t \) be 5mC levels on the M value scale for the \( p \)th CpG site and the \( i \)th individual. Using similar notations as above, we estimated the dispersion effect of age by the parameter \( \theta_p^\epsilon \) in the models,

\[ y_i^t = \mu^p + c_i^t \beta_p^c + (\text{SNP}_j^p \beta_p^\text{SNP} + \text{DateOfSampling}_j^p + \epsilon^p), \]

where

\[ \epsilon^p - \mathcal{N}(0, \sigma^2_p), \log_\text{10} \sigma_p = \tau_p + \text{Age} \theta_p^\epsilon. \]

We devised a hypothesis test for \( \theta \) by a likelihood ratio test comparing the model in Eq. (11), to a model with

\[ \epsilon^p - \mathcal{N}(0, \sigma^2_p), \log_\text{10} \sigma_p = \tau_p. \]

As a sensitivity analysis, we also fitted a model with

\[ \epsilon^p - \mathcal{N}(0, \sigma^2_p), \log_\text{10} \sigma_p = \tau_p + \text{Age} \theta_p^c + c_i^t \theta_p^c. \]
In this case, the hypothesis test for $\theta$ was done by comparing to a model with
\[ e_{p} \overset{\mathcal{N}}{\sim} (0, \sigma_{p}^{2}), \quad \log a_{i,p} = t^p + c_i^p \beta_p. \] (14)

These models were fitted with the gamlss R package\[^{83}\].

Local meQTL mapping analyses

Local meQTL mapping was performed using the MatrixEQTL R package\[^{8,9}\]. Association testing was performed 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, as predictors, the 15 immune cell proportions, a nonlinear age term encoded by 3 degrees-of-freedom (DoF) natural splines, sex, smoker status, ex-smoker status and CMV serostatus. 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\[^{54}\]. For the $i$th individual and the $p$th CpG site, let $y_{ip}$ be the measured 5mC levels on the M value scale, SNP_{pm}^{m,\text{m}}$ the minor allele count of the $m$th tested SNP for the CpG site and $f_{pm}^{m,\text{Age}}$ a nonlinear age term of natural splines. Moreover, let the vector $\mathbf{c}_i$ be measured proportions of the 15 immune cell-types for the $i$th individual, excluding neutrophils, and $\mathbf{\beta}_m^{\text{m}}$ the corresponding parameter vector. The additive allele effect of the SNP was estimated by the parameter $\mathbf{\beta}_m^{\text{m}}$ in the models,
\[ y_{ip} = \mu_{ip}^{m} + \text{SNP}_{ip}^{m} \beta_{\text{SNP}}^{m} + \text{Age}_{ip}^{m} \beta_{\text{Age}}^{m} + \text{CMV}_{ip}^{m} \beta_{\text{CMV}}^{m} + \text{Smoker}_{ip}^{m} \beta_{\text{Smoker}}^{m} + \text{PC1}_{ip}^{m} \beta_{\text{PC1}}^{m} + \text{PC2}_{ip}^{m} \beta_{\text{PC2}}^{m} + \epsilon_{ip}^{m}, \] (15)
where $\epsilon_{ip}^{m}$ is a symmetrical zero-mean distribution with constant variance.

Remote meQTL mapping analyses

Testing all possible associations between 644,517 CpG sites and 5,699,237 SNPs would require performing 3769 billion statistical tests. To reduce the multiple testing burden, remote meQTL mapping was conducted on a selection of 50,000 CpG sites with the highest residual variance in the model described in Eq. (15), but with $m$ indexing in this case only the most associated local SNP for the $p$th CpG 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. (15) but with $m$ now indexing remote SNPs for the $p$th CpG site. Both local and remote meQTL mapping tests were corrected for multiple testing by the Bonferroni adjustment.

Detection of independent remote meQTLs

We designed the following scheme to compute a set $\Phi$ of independently associated remote SNPs for each CpG site, where all such SNPs are associated with 5mC levels $y$ at the $p$th CpG site, conditional on the most associated local SNP and other SNPs in $\Phi$. Define $X_i$ to be the set of SNPs with a remote association to $y$ and let $x^0$ be the most associated significant local SNP, if it exists. The set $X_i$ typically includes several 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 $\Phi$ is constructed by elements of the final set $M$ that are associated with 5mC levels at the CpG site conditional on all the other elements in $M$. Intuitively, $\Phi$ consists of the most associated SNP in each LD block. The algorithm is given in pseudocode in Algorithm (1), where the condition $P^* = 0$ is determined by an $F$-test on the level $a = 10^{-6}$.

**Algorithm 1.** Forming a set of remote 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 $j = 1, 2, ...$

1. $P_j = \{ x \in X_j : y_{ip} = \mu_{ip}^{m} + x_i \beta_{ip}^{m} + \sum_{k=1}^{M_{j-1}} z_{ik} \beta_{ip}^{k} + \epsilon_{ip}^{m} \} \cap \{ 0, \sigma_{p}^{2} \}$

2. $P = \emptyset$

3. Exit

4. $X_j = M_1 \cup \{ x : x \text{ SNP with the smallest } P \text{-value in } P \}$

End

Cell-type-dependent effects of genetic and non-genetic factors on DNA methylation

To investigate whether the effects of a factor on DNA methylation depend on the proportion of myeloid cells in blood, we fitted models that included an interaction term between the factor of interest (i.e., age, sex, smoking status, CMV serostatus and genetic variants) and the proportion of myeloid cells, $\alpha$, defined as the sum of the proportions of cell-types from the myeloid lineage. With the same notations as above, but with $y^{c}$ being Smc levels on the $B$ value scale for the $p$th CpG site and the $i$th individual, we estimated the cell-type-dependent effects of non-genetic factors by fitting the models,
\[ y_{ip}^{c} = \mu^{c} + \text{Age}_{ip}^{c} + \text{CMV}_{ip}^{c} + \text{Smoker}_{ip}^{c} + \text{PC1}_{ip}^{c} + \text{PC2}_{ip}^{c} + \epsilon_{ip}^{c}, \] (16)
\[ y_{ip}^{c} = \mu^{c} + \text{Age}_{ip}^{c} + \text{CMV}_{ip}^{c} + \text{Smoker}_{ip}^{c} + \text{PC1}_{ip}^{c} + \text{PC2}_{ip}^{c} + \epsilon_{ip}^{c}, \] (17)

Inference in both cases was done by Wald tests with heteroscedasticity-consistent standard errors estimated by the sandwich R package\[^{17}\].

Detection of gene × environment interactions

We tested whether age, sex, CMV serostatus, smoking status or CRP levels could have a genotype-dependent effect on the DNA methylation. For the $i$th individual and the $p$th CpG site, let $y_{ip}^{e}$ be the 5mC levels on the $B$ value scale, SNP_{pk}^{m}, k = 1, \ldots, K, the minor allele counts of the significant local meQTL with the lowest $P$-value and the $K - 1$ independently associated remote meQTLs, and $\mathbf{c}_i$ the vector of 15 measured immune cell proportions with corresponding parameter vector $\mathbf{\beta}_m^{e}$. Interaction effects were estimated for each CpG site in the
model,
\[
E \left\{ y_i^p | \text{SNP}^{p,1}, \ldots, \text{SNP}^{p,K_p}, \text{Age}_i, \text{Woman}_i, \text{Smoker}_i, \text{CMV}_i \right\} \\
= \mu^p + \sum_{k=1}^{K_p} \text{SNP}^{p,k} \beta_{p,k}^{y_i} + \text{PC}_1^p \beta_{p,\text{PC}_1}^c + \text{PC}_2 \beta_{p,\text{PC}_2}^c + \text{Age}_i \beta_{p,\text{Age}}^c \\
+ \text{Woman}_i \beta_{p,\text{Woman}}^c + \text{Smoker}_i \beta_{p,\text{Smoker}}^c + \text{CMV}_i \beta_{p,\text{CMV}}^c \\
+ \sum_{k=1}^{K_p} \text{SNP}^{p,k} \left( \text{Age}_i \beta_{c,\text{Age}}^c + \text{Woman}_i \beta_{c,\text{Woman}}^c + \text{Smoker}_i \beta_{c,\text{Smoker}}^c + \text{CMV}_i \beta_{c,\text{CMV}}^c \right)
\]
(18)

We investigated effects of CRP levels in a separate model that simply added a log-transformed CRP term to Eq. (18). Inference was done by Wald tests with heteroscedasticity-consistent standard errors estimated by the sandwich R package \cite{20}.

**Estimation of proportions of explained 5mC variance**

According to our analyses, 5mC levels in the healthy population are mainly associated with local genetic variation, blood cell composition, age, sex, smoking, CMV infection and CRP levels. We grouped these variables into four categories: genetic, cell composition, intrinsic (age and sex) and exposures (smoking, CMV infection and CRP levels). For the \( p \)-th CpG site and the \( i \)-th individual, we collected observations of the minor allele count for the most associated local SNP in \( x_i^{p,k} \), the proportions of the 15 cell types, excluding neutrophils, in the vector \( x_i^c \), intrinsic factors (sex and natural spline expanded values of age) in the vector \( x_i^a \), and exposures (smoking status, CMV serostatus and log-transformed CRP levels) in the vector \( x_i^e \), with corresponding parameters \( \beta_e^c, \beta_c^e, \beta_m^c \) and \( \beta_e^c \). We interpret here log-transformed CRP levels as a proxy measure of the exposure of chronic low-grade inflammation. For each CpG site, we define linear predictor terms by

\[
f_p^e(x_i^{p,k}) = x_i^{p,k} \beta_e^c, \\
f_p^e(x_i^c) = (x_i^c)^\top \beta_c^e, \\
f_p^e(x_i^a) = (x_i^a)^\top \beta_m^c, \\
f_p^e(x_i^e) = (x_i^e)^\top \beta_e^c
\]
(19-22)

These functions vary in their degrees of freedom, 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 models on samples from \( I_1 \), and evaluating the prediction accuracy on samples from \( I_2 \).

Let \( y_i^e \) be 5mC levels for the \( p \)-th CpG site on the \( \beta \) value scale. Take cell composition as an example. To compute the total effect of cell composition on 5mC levels at the CpG site, we first fit a model with individuals in \( I_1 \),

\[
y_i^{p,c} = \mu^p + (x_i^c)^\top \beta_c^e, \quad i \in I_1
\]
(23)

with parameters \( \beta_c^e \) and \( \mu^p \) estimated by least squares. We then define the total effect size to be the squared correlation between the observations and the out-of-sample predictions in individuals in \( I_2 \),

\[
\left( \hat{R}_{Tot}^2 \right) = \text{corr} \left\{ y_i^{p,c}, \hat{y}_{i}^{p,c} \right\}^2, \quad j \in I_2.
\]
(24)

Total effects for the other predictor groups were defined analogously.

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. Take the exposures group as an example, the direct effect was computed by

\[
\left( \hat{R}_{Tot}^2 \right) = \left( \hat{R}_{Tot+c}^2 \right) - \left( \hat{R}_{Tot}^2 \right),
\]
(25)

where \( \hat{R}_{Tot+c}^2 \) is the total effect of the sum of the predictor terms for exposures and cell composition.

\[
f_{c+e} = f_p^e(x_i^c) + f_p^e(x_i^e).
\]
(26)

To mitigate the impact of sampling on estimates of total and direct effects, we did four independent repeats of five-fold cross-validation and averaged effect sizes across all 20 samples. To have an unbiased estimation of the out-of-sample explained variance, we re-fit 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:

With equal probability, assign an integer between 1 and 5 to all individuals.

For \( k = 1, \ldots, 5 \)

Index individuals assigned \( k \) as \( I_k \), the others are indexed as \( I_{-k} \).

Select SNP for the predictor \( f_p^e \) by performing a local meQTL mapping on individuals in \( I_k \).

For predictor \( f_p^e \), \( I_k \in \{ I_p^f \cap f^p_{in}, f^p_{out} \} \)

Estimate \( \mu^p, \beta_e^c \) with \( I_{-k} \).

Compute \( \left( \hat{R}_{Tot}^2 \right) \) by Eq. (24) with \( I_2 = I_k \).

The scheme to sample the direct effects is analogous. Finally, we estimated an effect size for interactions between the local SNP and non-genetic factors for each CpG site. It was computed, similarly to Eq. (25), as the added out-of-sample proportion of variance explained by the regression function,

\[
f_{\text{int}}^p (\text{SNP}^p, \text{Age}_i, \text{Woman}_i, \text{CMV}_i, \text{ExSmoker}_i, \text{Smoker}_i, \text{CRP}_i) \\
= \mu^p + \text{SNP}^p \beta_{\text{SNP}}^e + \text{Age}_i \beta_{\text{Age}}^e + \text{Woman}_i \beta_{\text{Woman}}^e + \text{CMV}_i \beta_{\text{CMV}}^e + \text{ExSmoker}_i \beta_{\text{ExSmoker}}^e + \text{Smoker}_i \beta_{\text{Smoker}}^e + \log(\text{CRP}_i) \beta_{\text{CRP}}^e + \text{SNP}^p \langle \text{Age}, \beta_{\text{Age}}^e, \text{Woman}, \beta_{\text{Woman}}^e, \text{CMV}, \beta_{\text{CMV}}^e, \text{ExSmoker}, \beta_{\text{ExSmoker}}^e, \text{Smoker}, \beta_{\text{Smoker}}^e, \log(\text{CRP}) \beta_{\text{CRP}}^e \rangle
\]
(27)

compared to the same regression function without interaction terms,

\[
f_{\text{main}}^p (\text{SNP}^p, \text{Age}_i, \text{Woman}_i, \text{CMV}_i, \text{ExSmoker}_i, \text{Smoker}_i, \text{CRP}_i) \\
= \mu^p + \text{SNP}^p \beta_{\text{SNP}}^e + \text{Age}_i \beta_{\text{Age}}^e + \text{Woman}_i \beta_{\text{Woman}}^e + \text{CMV}_i \beta_{\text{CMV}}^e + \text{ExSmoker}_i \beta_{\text{ExSmoker}}^e + \text{Smoker}_i \beta_{\text{Smoker}}^e + \log(\text{CRP}_i) \beta_{\text{CRP}}^e
\]
(28)

**Biological annotations**

Information about the position, closest gene and CpG density of each CpG site was obtained from the Illumina EPIC array manifest v1.0 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 ROADMp Epigenomics consortium \cite{21}. We used
peripheral blood mononuclear cells (PBMCs) as reference. 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 database15,16, including a total of 1165 TFs. Binding sites include both direct and indirect binding. Enrichment analyses were performed by creating simple two-way tables for each target set and each annotation (i.e., chromatin states, CpG density, transcription factor binding site), and then performing Fisher's exact test. Gene ontology enrichments were computed with the goemethod function in the missMethyl R package119.

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 LD (set here to $r^2 > 0.6$) with a disease or trait-associated variant ($P$-value $< 5 \times 10^{-8}$). EBI-NHGRI Catalog of GWAS hits version e100 r2021-01. 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)120.

Reporting summary
Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The Infinitium MethylationEPIC raw and processed data generated in this study22 have been deposited in the Institut Pasteur data repository, OWY, which can be accessed via the following link: https://doi.org/10.4880/owy.F83s6I042. All association statistics obtained in this study (i.e., the 141 EWAS and interaction models, local meQTL mapping) can be explored and downloaded from the web browser http://mimeth.pasteur.fr/. The SNP array data can be accessed in the European Genome-Phenome Archive (EGA) with the accession code EGAS0000102460. All Milieu Intérieur datasets can be accessed by submitting a data access request to milieuinterieurdac@pasteur.fr, the Milieu Intérieur data access committee (DAC). The DAC informs all the submitting a data access request to milieuinterieurdac@pasteur.fr, the Genome-Phenome Archive (EGA) with the accession code EGAS0000102460. All Milieu Intérieur datasets can be accessed by submitting a data access request to milieuinterieurdac@pasteur.fr, the Milieu Intérieur data access committee (DAC). The DAC informs all the participants of the data access request and grants data access if the request is consistent with the informed consent signed by the participants. In particular, research on Milieu Intérieur datasets is restricted to research on the genetic and environmental determinants of human variation in immune responses. Data access is typically granted two months after request submission.

Code availability
All the code supporting the current study, including the CMV estimation model, has been uploaded to GitHub:231 https://github.com/JacobBergstedt/MIMETH.

References
1. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. Nature 571, 489–499 (2019).
2. Michalak, E. M., Burr, M. L., Bannister, A. J. & Dawson, M. A. The roles of DNA, RNA and histone methylation in ageing and cancer. Nat. Rev. Mol. Cell Biol. 20, 573–589 (2019).
3. Martin, E. M. & Fry, R. C. Environmental influences on the epigenome: exposure-associated DNA methylation in human populations. Annu. Rev. Public Health 39, 309–333 (2018).
4. Karlsson Linner, R. et al. An epigenome-wide association study meta-analysis of educational attainment. Mol. Psychiatry 22, 1680–1690 (2017).
5. Lam, L. L. et al. Factors underlying variable DNA methylation in a human community cohort. Proc. Natl Acad. Sci. USA 109, 17253–17260 (2012).
6. Stringhini, S. et al. Life-course socioeconomic status and DNA methylation of genes regulating inflammation. Int. J. Epidemiol. 44, 1320–1330 (2015).
7. Bush, N. R. et al. The biological embedding of early-life socioeconomic status and family adversity in children’s genome-wide DNA methylation. Epigenetics 10, 1445–1461 (2018).
8. Hwang, J. Y., Aromolaran, K. A. & Zukin, R. S. The emerging field of epigenetics in neurodegeneration and neuroprotection. Nat. Rev. Neurosci. 18, 347–361 (2017).
9. Mazzone, R. et al. The emerging role of epigenetics in human autoimmune disorders. Clin. Epigenet. 11, 34 (2019).
10. Ling, C. & Ronn, T. Epigenetics in human obesity and type 2 diabetes. Cell Metab. 29, 1028–1044 (2019).
11. van der Harst, P., de Windt, L. J. & Chambers, J. C. Translational perspective on epigenetics in cardiovascular disease. J. Am. Coll. Cardiol. 70, 590–606 (2017).
12. Wild, C. F. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol. Biomark. Prev. 14, 1847–1850 (2005).
13. Berdasco, M. & Esteller, M. Clinical epigenetics: seizing opportunities for translation. Nat. Rev. Genet. 20, 109–127 (2019).
14. Farñik, M. et al. DNA methylation dynamics of human hematopoietic stem cell differentiation. Cell Stem Cell 19, 808–822 (2016).
15. Roadmap Epigenomics Consortium, et al. Integrative analysis of 111 reference human epigenomes. Nature 516, 317–330 (2015).
16. Liu, Y. et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat. Biotechnol. 31, 142–147 (2013).
17. Lappalainen, T. & Greally, J. M. Associating cellular epigenetic models with human phenotypes. Nat. Rev. Genet. 18, 441–451 (2017).
18. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinforma. 13, 96 (2012).
19. Teschendorff, A. E., Breeze, C. E., Zheng, S. C. & Beck, S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC Bioinforma. 18, 105 (2017).
20. Lemire, M. et al. Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. Nat. Commun. 6, 6326 (2015).
21. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat. Genet. 49, 131–138 (2017).
22. Villicana, S. & Bell, J. T. Genetic impacts on DNA methylation: research findings and future perspectives. Genome Biol. 22, 127 (2021).
23. Min, J. L. et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat. Genet. 53, 1311–1321 (2021).
24. Thomas, S. et al. The Milieu Intérieur study - an integrative approach for study of human immunological variance. Clin. Immunol. 157, 277–293 (2015).
25. Patin, E. et al. Natural variation in the parameters of innate immune cells is preferentially driven by genetic factors. Nat. Immunol. 19, 302–314 (2018).
26. Houseman, E. A., Kelsey, K. T., Wiencke, J. K. & Marsit, C. J. Cell-composition effects in the analysis of DNA methylation array data: a mathematical perspective. BMC Bioinforma. 16, 95 (2015).
44. Torti, N., Walton, S. M., Murphy, K. M. & Oxenius, A. Batf3 transcription factor-dependent DC subsets in murine CMV infection: differential impact on T-cell priming and memory in older individuals. *Aging Cell* **12**, 381–387 (2013).
45. Savva, G. M. et al. Cytomegalovirus infection is associated with increased mortality in the older population. *Aging Cell* **12**, 381–387 (2013).
46. Chen, S. et al. Associations of cytomegalovirus infection with all-cause and cardiovascular mortality in multiple observational cohort studies of older adults. *J. Infect. Dis.* **223**, 238–246 (2021).
47. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
48. Heyn, H. et al. Distinct DNA methylation of newborns and centenarians. *Proc. Natl. Acad. Sci. USA* **109**, 10522–10527 (2012).
49. Johansson, A., Enroth, S. & Gyllensten, U. Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS ONE* **8**, e67378 (2013).
50. Wang, Y. et al. Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. *Epigenetics* **13**, 975–987 (2018).
51. Jones, M. J., Goodman, S. J. & Kobor, M. S. DNA methylation and healthy human aging. *Aging Cell* **14**, 924–932 (2015).
52. Lachmann, R. et al. Cytomegalovirus (CMV) seroprevalence in the adult population of Germany. *PLoS ONE* **13**, e0200267 (2018).
53. McCartney, D. L. et al. An epigenome-wide association study of sex-specific chronological ageing. *Genome Med.* **12**, 1 (2019).
54. Zhu, T., Zheng, S. C., Paul, D. S., Horvath, S. & Teschendorff, A. E. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. *Aging* **10**, 3541–3557 (2018).
55. Bracken, A. P. et al. The Polycrom group proteins bind throughout the INK4AR locus in aged and disassociated in senescent cells. *Genes Dev.* **21**, 525–530 (2007).
56. Siebold, A. P. et al. Polycrom Repressive Complex 2 and Trithorax modulate Drosophila longevity and stress resistance. *Proc. Natl. Acad. Sci. USA* **107**, 168–174 (2010).
57. Boyer, L. A. et al. Polycrom complexes repress developmental regulators in murine embryonic stem cells. *Nature* **441**, 349–353 (2006).
58. Dozmorov, M. G. Polycomb repressive complex 2 epigenomic signature defines age-associated hypermethylation and gene expression changes. *Epigenetics* **10**, 484–495 (2015).
59. Fraga, M. F. et al. Epigenetic differences arise during the lifetime of monozygotic twins. *Proc. Natl. Acad. Sci. USA* **102**, 10604–10609 (2005).
60. Slyker, R. et al. Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17**, 191 (2016).
61. Zhang, Q. et al. Genotype effects contribute to variation in longitudinal methylome patterns in older people. *Genome Med.* **7**, 77 (2015).
62. Nikolich-Zugich, J. The twilight of immunity: emerging concepts in healthy aging. *Nat. Immunol.* **19**, 10–19 (2018).
63. Singmann, P. et al. Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenet. Chromatin* **8**, 43 (2015).
64. Yousif, P. et al. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics* **16**, 911 (2015).
65. Gatev, E. et al. Autosomal sex-associated co-methylated regions predict biological sex from DNA methylation. *Nucleic Acids Res.* **49**, 9097–9116 (2021).
66. Cortes-Saez, A. et al. Updating dual-specificity tyrosine-phosphorylation-regulated kinase 2 (DYRK2): molecular basis, functions and role in diseases. *Cell Mol. Life Sci.* **77**, 4747–4763 (2020).
67. Yusipov, I. et al. Age-related DNA methylation changes are sex-specific: a comprehensive assessment. *Aging* **12**, 24057–24080 (2020).
68. Karlsson Linner, R. et al. 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 (2019).
69. Lee, J. J. et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. Nat. Genet. 50, 112–112 (2018).

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

71. Feinberg, A. P. The role of epigenetics in human disease prevention and mitigation. N. Engl. J. Med. 378, 1323–1334 (2018).

72. O’Geen, H. et al. Genome-wide analysis of KAPI binding suggests autoregulation of KRAB-ZNFs. PLoS Genet. 3, e89 (2007).

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

74. Quenneville, S. et al. 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 (2012).

75. Hawe, J. S. et al. Genetic variation influencing DNA methylation provides insights into molecular mechanisms regulating genomic function. Nat. Genet. 54, 18–29 (2022).

76. Emilsson, V. et al. Co-regulatory networks of human serum proteins link genetics to disease. Science 361, 769–773 (2018).

77. Koller, C. J. et al. BACE2 expression increases in human neurodegenerative disease. Am. J. Pathol. 180, 337–350 (2012).

78. Granada, M. et al. A genome-wide association study of plasma total IgE concentrations in the Framingham Heart Study. J. Allergy Clin. Immunol. 129, 840–845 e21 (2012).

79. Beerman, I. et al. Proliferation-dependent alterations of the DNA methylation landscape underlie hematopoietic stem cell aging. Cell Stem Cell 12, 413–425 (2013).

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

81. Li, Y. et al. Genome-wide analyses reveal a role of Polycamb in promoting hypomethylation of DNA methylation valleys. Genome Biol. 19, 18 (2018).

82. Kim, J. Y., Tavare, S. & Shibata, D. Counting human somatic cell replications: methylation mirrors endometrial stem cell divisions. Proc. Natl Acad. Sci. USA 102, 17739–17744 (2005).

83. Yang, Z. et al. Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. 17, 205 (2016).

84. Zhou, W. et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat. Genet. 50, 591–602 (2018).

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

86. Bell, J. T. et al. Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet. 8, e1002629 (2012).

87. Grundberg, E. et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat. Genet. 44, 1084–1089 (2012).

88. van Dongen, J. et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat. Commun. 7, 11115 (2016).

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

90. Li, S. et al. Causes of blood methylocnic variation for middle-aged women measured by the HumanMethylation450 array. Epigenetics 12, 973–981 (2017).

91. Boyce, W. T. & Kobor, M. S. Development and the epigenome: the ‘synapse’ of gene-environment interplay. Dev. Sci. 18, 1–23 (2015).

92. Fleiss, J. L. Design and Analysis of Clinical Experiments (Wiley, 2011).

93. Czamara, D. et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nat. Commun. 10, 2548 (2019).

94. Teh, A. L. et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 24, 1064–1074 (2014).

95. Bergstedt, J., Urrutia, A., Albert, M. L., Quintana-Murci, L. & Patin, E. Accurate prediction of cell composition, age, smoking consumption and infection serostatus based on blood DNA methylation profiles. Preprint at bioRxiv https://doi.org/10.1101/456996 (2018).

96. Furman, D. et al. Chronic inflammation in the etiology of disease across the life span. Nat. Med. 25, 1822–1832 (2019).

97. Mauvais-Jarvis, F. et al. Sex and gender: modifiers of health, disease, and medicine. Lancet 396, 565–582 (2020).

98. Nicoli, T. & Partridge, L. Ageing as a risk factor for disease. Curr. Biol. 22, R741–R752 (2012).

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

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

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

102. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear-mixed effects models using lme4. J. Stat. Soft. 67, 1–48 (2015).

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

104. Price, M. E. et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenet. Chromatin 6, 4 (2013).

105. Auton, A. et al. A global reference for human genetic variation. Nature 526, 68–74 (2015).

106. Stekhoven, D. J. & Bühlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 28, 112–118 (2012).

107. Manichaikul, A. et al. Robust relationship inference in genome-wide association studies. Bioinformatics 26, 2867–2873 (2010).

108. Pawlowsky-Glahn, V., José Egozcue, J. & Tolosana-Delgado, R. Modern compositional data analysis. Inter J. Epidemicol. 49, 1307–1313 (2020).

109. Templ, M., Hron, K. & Filzmoser, P. robCompositions: an R-package for robust statistical analysis of compositional data (John Wiley and Sons, 2011).

110. Halekoh, U. & Hejsgaard, S. A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – The R Package pbkrtest. 59, 32 (2014).

111. Salas, L. A. et al. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. Genome Biol. 19, 64 (2018).

112. Zou, H. & Hastie, T. Regularization and variable selection via the lasso. J. R. Stat. Soc.: Ser. B 67, 301–320 (2005).

113. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. J. Stat. Soft 33, 1–22 (2010).
115. Stasinopoulos, M. D., Rigby, R. A. & Bastiani, F. D. GAMLSS: a
distributional regression approach. Stat. Model. 18, 248–273 (2018).
116. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large
matrix operations. Bioinformatics 28, 1353–1358 (2012).
117. Zeileis, A., Köll, S. & Graham, N. Various versatile variances: an
object-oriented implementation of clustered covariances in R. J.
Stat. Softw. 95, 36 (2020).
118. Cheney, J. et al. 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 (2020).
119. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R pack-
geage for analyzing data from Illumina’s HumanMethylation450
platform. Bioinformatics 32, 286–288 (2016).
120. Chang, C. C. et al. Second-generation PLINK: rising to the chal-
genle of larger and richer datasets. GigaScience 4, 7 (2015).
121. Bergstedt, J. The immune factors driving DNA methylation varia-
tion in human blood. GitHub, https://doi.org/10.5281/zenodo.
7016878 (2022).

Acknowledgements
We thank Sarah Merrill, Nicole Gladish, Violaine Saint-André, Lucas
Husquin and the Milieu Intérieur scientific advisory board for comments
and helpful discussions. We acknowledge the help of the HPC Core
Facility of Institut Pasteur for this work. This research was enabled, in
part, by the use of the FlowSorted.BloodExtended.EPIC R package
developed at Dartmouth College, which software is subject to the
licensing terms made available by Dartmouth Technology Transfer and
which software is provided “as is” with no warranties whatsoever. 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 devel-
oped 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 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 interpreta-
tion. J.B., E.P., and L.Q.-M. wrote the manuscript. All authors discussed
the results and contributed to the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary
material available at https://doi.org/10.1038/s41467-022-33511-6.

Correspondence and requests for materials should be addressed to
Jacob Bergstedt, Etienne Patin or Lluís Quintana-Murci.

Peer review information Nature Communications thanks the anon-
ymous reviewers for their contribution to the peer review of this
work. Peer reviewer reports are available.

Reprints and permission information is available at
http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jur-
isdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons
Attribution 4.0 International License, which permits use, sharing,
adaptation, distribution and reproduction in any medium or format, as
long as you give appropriate credit to the original author(s) and the
source, provide a link to the Creative Commons license, and indicate if
changes were made. The images or other third party material in this
article are included in the article’s Creative Commons license, unless
indicated otherwise in a credit line to the material. If material is not
included in the article’s Creative Commons license and your intended
use is not permitted by statutory regulation or exceeds the permitted
use, you will need to obtain permission directly from the copyright
holder. To view a copy of this license, visit http://creativecommons.org/
licenses/by/4.0/.

© The Author(s) 2022

Milieu Intérieur Consortium
Laurent Abel, Andrés Alcover, Hugues Aschard, Philippe Bouso, Nollaig Bourke, Petter Brodin, Pierre Bruhns, Nadine Cerf-Bensussan, Ana Cumanco, Christophe d’Enfert, Ludovic Deriano, Marie-Agnès Dillies, James Di Santo, Françoise Dromer, Gérard Eben, Jost Enninga, Jacques Fellay, Ivo Gomperts-Boneca, Milena Hasan, Gunilla Karlsson Hedestam, Serge Hercberg, Molly A. Ingersoll, Olivier Lantz, Rose Anne Kennedy, Mickaël Ménager, Frédéric Michel, Hugo Mouquet, Cliona O’Farrell, Etienne Patin, Sandra Pellegrini, Antonio Rausell, Frédéric Rieux-Laucat, Lars Rogge, Magnus Fontes, Anavaj Sakuntabhai, Olivier Schwartz, Benoît Schwikowski, Spencer Shorte, Frédéric Tangy, Antoine Toubert, Mathilde Touvier, Marie-Noëlle Ungeheuer, Christophe Zimmer, Matthew L. Albert, Darragh Duffy & Lluís Quintana-Murci.

8Imagine Institute, University Paris Cité, Necker Hospital for Sick Children, INSERM UMR 1163, Laboratory of Human Genetics of Infectious Diseases, Paris, France. 9Institut Pasteur, Université Paris Cité, INSERM-U1224, Unité Biologie Cellulaire des Lymphocytes, Ligue Nationale Contre le Cancer, Équipe Labellisée Ligue, 2018 Paris, France. 10Institut Pasteur, Université Paris Cité, Department of Computational Biology, Paris, France. 11Institut Pasteur, Université Paris Cité, INSERM UMR1222, Unit of Antibodies in Therapy and Pathology, Paris, France. 12Institut Imagine, Université Paris Cité, INSERM UMR1163, Laboratory Intestinal Immunity, Paris, France. 13Institut Pasteur, Université Paris Cité, INSERM UMR1163, Laboratory of Human Genetics of Infectious Diseases, Paris, France. 14Department of Immunology and Inflammation, Imperial College London, London, UK. 15Department of Women’s and Children’s Health, Karolinska Institutet, Stockholm, Sweden. 16Institut Pasteur, Université Paris Cité, INSERM UMR1222, Unit of Antibodies in Therapy and Pathology, Paris, France. 17Institut Imagine, Université Paris Cité, INSERM UMR1163, Laboratory Intestinal Immunity, Paris, France.
