bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Title

Epigenome-wide change and variation in DNA methylation from birth to late adolescence

Authors

Rosa H. Mulder[1,2,3], Alexander Neumann[1,2,4]*, Charlotte A. M. Cecil[1,5,6]*, Esther Walton[7,8], Lotte C. Houtepen[7], Andrew J. Simpkin[7,9], Jolien Rijlaarsdam[1,2], Bastiaan T. Heijmans[10], Tom R. Gaunt[7], Janine F. Felix[2,11], Vincent W. V. Jaddoe[2,11], Marian J. Bakermans-Kranenburg[12], Henning Tiemeier[1,13], Caroline L. Relton[7], Marinus H. van IJzendoorn[14,15], and Matthew Suderman[7]

*both authors contributed equally to this manuscript

Affiliations

1. Department of Child and Adolescent Psychiatry/Psychology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

2. Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

3. Institute of Education and Child Studies, Leiden University, Leiden, the Netherlands

4. Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, Qc, Canada

5. Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

6. Department of Psychology, Institute of Psychology, Psychiatry & Neuroscience, King's College London, London, UK

7. MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

8. Department of Psychology, University of Bath, Bath, UK

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. School of Mathematics, Statistics and Applied Mathematics, National University of Ireland, Galway, Ireland

10. Molecular Epidemiology, Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

11. Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands

12. Clinical Child & Family Studies, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

13. Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, Boston USA

14. Department of Psychology, Education and Child Studies, Erasmus University Rotterdam, Rotterdam, the Netherlands

15. School of Clinical Medicine, University of Cambridge, Cambridge, UK

Corresponding author

Matthew Suderman

matthew.suderman@bristol.ac.uk

Tel: +44 (0)117 331 0090

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

DNA methylation (DNAm) is known to play a pivotal role in childhood health and development, but a comprehensive characterization of genome-wide DNAm trajectories across this age period is currently lacking. We have therefore performed a series of epigenome-wide association studies in 5,019 blood samples collected at multiple time-points from birth to late adolescence from 2,348 participants of two large independent cohorts. DNAm profiles of autosomal CpG sites (CpGs) were generated using the Illumina Infinium HumanMethylation450 BeadChip. Change over time was widespread, observed at over one-half (53%) of CpGs. In most cases DNAm was decreasing (36% of CpGs). Inter-individual variation in linear trajectories was similarly widespread (27% of CpGs). Evidence for nonlinear change and inter-individual variation in nonlinear trajectories was somewhat less common (11% and 8% of CpGs, respectively). Very little inter-individual variation in change was explained by sex differences (0.4% of CpGs) even though sex-specific DNAm was observed at 5% of CpGs. DNAm trajectories were distributed non-randomly across the genome. For example, CpGs with decreasing DNAm were enriched in gene bodies and enhancers and were annotated to genes enriched in immune-developmental functions. By contrast, CpGs with increasing DNAm were enriched in promoter regions and annotated to genes enriched in neurodevelopmental functions. These findings depict a methylome undergoing widespread and often nonlinear change throughout childhood. They support a developmental role for DNA methylation that extends beyond birth into late adolescence and has implications for understanding life-long health and disease. DNAm trajectories can be visualized at http://epidelta.mrcieu.ac.uk.

Key words

EWAS, longitudinal, ageing, DNA methylation, change, childhood, trajectories, inter-individual variation, Generation R, ALSPAC

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Introduction

DNA methylation (DNAm), an epigenetic process whereby DNA is modified by the addition of methyl groups, has gained increasing attention over the past few decades, due to its pivotal role in development. *In utero*, DNAm is involved in a range of essential processes including cell differentiation[1-3], X-chromosome inactivation[4] and fetal growth[5]. Its role extends well beyond birth, e.g. by maintaining cell type identity and genome stability[6-8], responding to environmental exposures[9-11], and its involvement in immune[12] and neural development[13]. Since it is influenced by *both* genetic and environmental factors[14,15], DNAm has also emerged as a key mechanism of interest for understanding the gene-environmental interplay in normal ageing and disease development.

Numerous studies have identified strong associations between DNAm and age. While most have relied on cross-sectional data[16-18], but a few have utilized longitudinal measurements of DNAm within individuals[19-23]. Longitudinal measurements allow one to distinguish intra-individual change from inter-individual differences in change, thereby greatly improving the power to detect change over time and to identify differences between individuals[24]. Identifying and characterizing CpGs for which DNAm changes differently over time between individuals (i.e. inter-individual variation in change) is a necessary step in identifying genetic and environmental influences on the methylome as well as their potential impact on health outcomes[25]. Moreover, longitudinal designs facilitate the study of nonlinear trajectories[26,27], which might help to identify sensitive periods for DNAm change in development. To date, the largest epigenome-wide longitudinal study on DNAm included 385 elderly individuals who were followed up to five times over a maximum period of 18 years, identifying DNAm change at 1,316 CpG (Cytosine-phosphate-Guanine) sites[19] and inter-individual variation at change at 570 CpGs[20]. Yet, little is known about DNAm trajectories across early development, as existing studies in *childhood* DNAm typically have been limited by small sample sizes[21,23], short time-periods[22,28] or focused on specific CpGs in relation to maternal smoking[29], birthweight[30], or maternal BMI[31].

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In the current study, we aim to provide a benchmark of typical epigenome-wide age-related DNAm trajectories within individuals, spanning the first two decades of life. This study combines repeated measurements of DNAm at nearly half a million CpG sites across the genome from two large population-based cohorts, the Generation R Study and Avon Longitudinal Study of Parents and Children (ALSPAC), to form one integrated dataset with four time-points of measurement. In a series of three epigenome-wide mixed model analyses we study linear (Model 1), nonlinear (Model 2) and sex-related (Model 3) trajectories of change across development. Further, we aim to identify CpGs for which trajectories vary between individuals (Model 1 and 2). Results are interpreted in the context of CpG location and biological pathways. The key findings are discussed here, full results per CpG can be freely accessed and visualized at http://epidelta.mrcieu.ac.uk/ [*note to reviewers: this is a demo website for now*].

Results

Cross-cohort comparability

Sample characteristics of 1,399 Generation R participants (total DNAm samples=2,333) and of 949 ALSPAC participants (total DNAm samples=2,686; Figure 1) are provided in Supplementary Table 1. After the DNAm datasets of the two cohorts underwent joint functional normalization (see Supplementary Figure 1 for distributions of mean DNAm levels), within-cohort stability of DNAm at birth and 6 or 7 years (in Generation R and ALSPAC, respectively) was compared. Stability of DNAm at individual CpG sites (437,864 autosomal sites) was estimated in three ways: relative concordance using Spearman correlations between time points, absolute concordance using intraclass correlations between time points (children with data for both time points: *n* Generation R=476, *n* ALSPAC=826), and change over time using change estimates from a linear mixed model (Model 1, online Methods) applied within each cohort (children with data for at least one of the two time-points: *n* Generation R=1,394, *n* ALSPAC=944). Estimates of all stability measures for both cohorts are depicted in Figure 2. Next, agreement of these stability estimates between the two cohorts was estimated with the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Spearman (ρ) or Pearson (*r*) correlation (depending on normality of the data) across all CpGs, between the datasets. The Spearman correlation of the relative concordance was ρ=0.62, the Pearson correlation of the absolute concordance was ρ=0.60, and the Pearson correlation of the change estimates was *r*=0.86, indicating strong agreement between datasets. Based on these results the two datasets were joined to form one set with four different time-points of DNAm (birth, age 6/7 years, 10 years, 17 years).

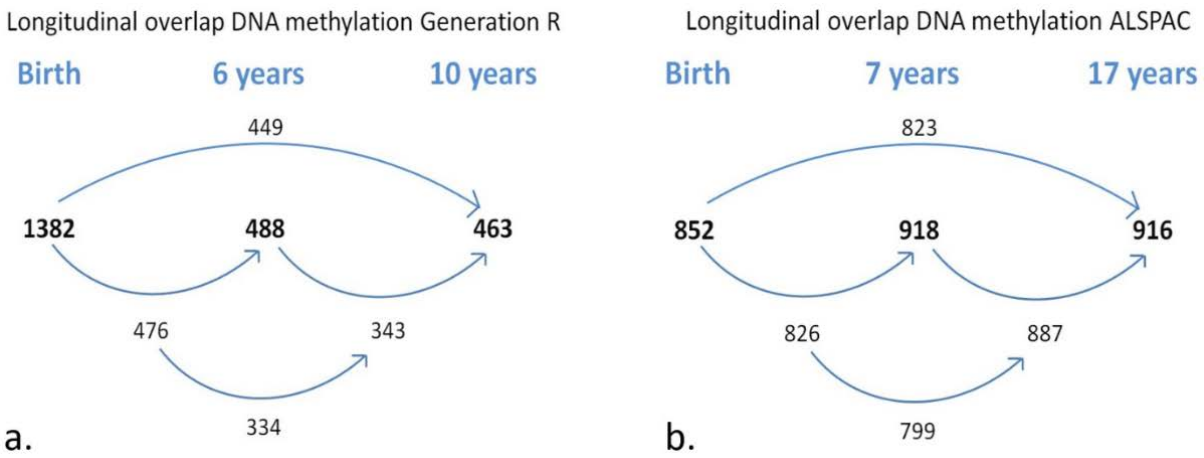
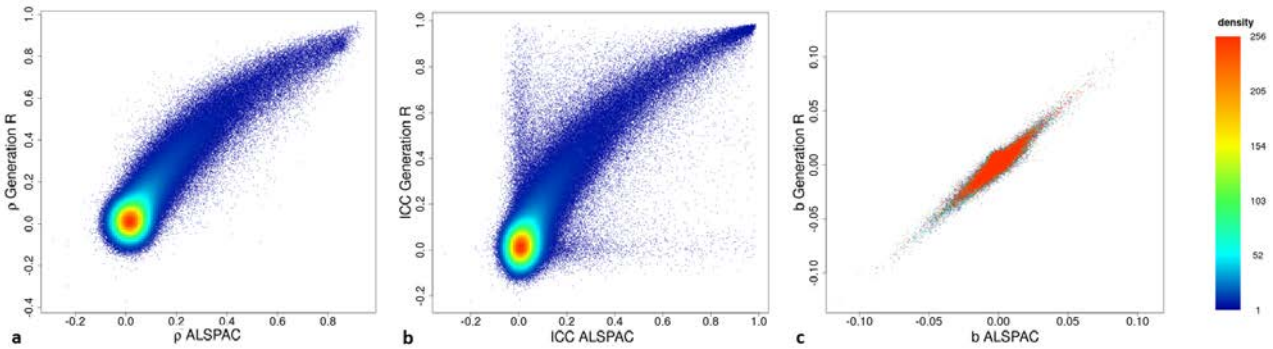


Figure 1. Longitudinal sample sizes. Sample sizes for (a) Generation R (N total children=1,399, N total DNAm samples=2,333); and (b) ALSPAC (N total children=949, N total DNAm samples=2,686). Bolded numbers represent total sample size at each time-point; non-bolded number refer to overlapping samples between time-points.



bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. Scatterplots of within-cohort stability of DNA methylation. Showing (a) Spearman correlations, (b) intraclass correlation coefficients and (c) change estimates from birth to 6/7 years per CpG for Generation R and ALSPAC.

Linear DNAm change from birth to early adulthood

Estimates of overall change in DNAm from birth to early adolescence (Model 1; see online Methods) indicated linear change at 51.6% of CpGs at a Bonferroni-corrected threshold ($P<1x10^{-07}$) (Figure 3a and 3b). Specifically, DNAm decreased over time at 35.5% of all CpGs and increased at 16.0%. The mode intercept indicated that the decreasing CpGs were 88% methylated at birth (Figure 4). DNAm levels for increasing CpGs typically started at 5%.

The mode estimate DNAm change was $b=-9.24x10^{-04}$ (with corresponding mode $SE=6.85x10^{-05}$), indicating an overall 0.09% DNAm decrease per year at a typical CpG site. This translates into a 1.66% decrease in DNAm over the course of 18 years. An example of a CpG site with a typical change in DNAm is depicted in Figure 3a. The largest observed absolute change in DNAm was $b=-3.47x10^{-02}$ ($SE=3.65x10^{-04}$, $P<9.88x10^{-324}$), indicating an overall DNAm decrease of 62.5% over 18 years (Figure 3b). Only twenty-two CpGs showed an absolute change >50% over the course of 18 years (Supplementary Table 2). From this it follows that typically in (cord-/peripheral) blood tissue, DNAm levels for CpGs do not change from a fully unmethylated to fully methylated state, or vice versa, over the course of 18 years.

Further, we observed substantial inter-individual variation in linear DNAm changes over time at 27.4% of all CpGs (i.e. random slope variance was greater than zero at Bonferroni-corrected threshold $P<1x10^{-07}$; Figure 3c). On average, this variation accounted for 2.7% (SD=1.5%) of all estimated inter-individual variation (for intercept, age, batch, and residual) at these CpGs. At 17.3% of all CpGs, we observed both change and inter-individual variation in change.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

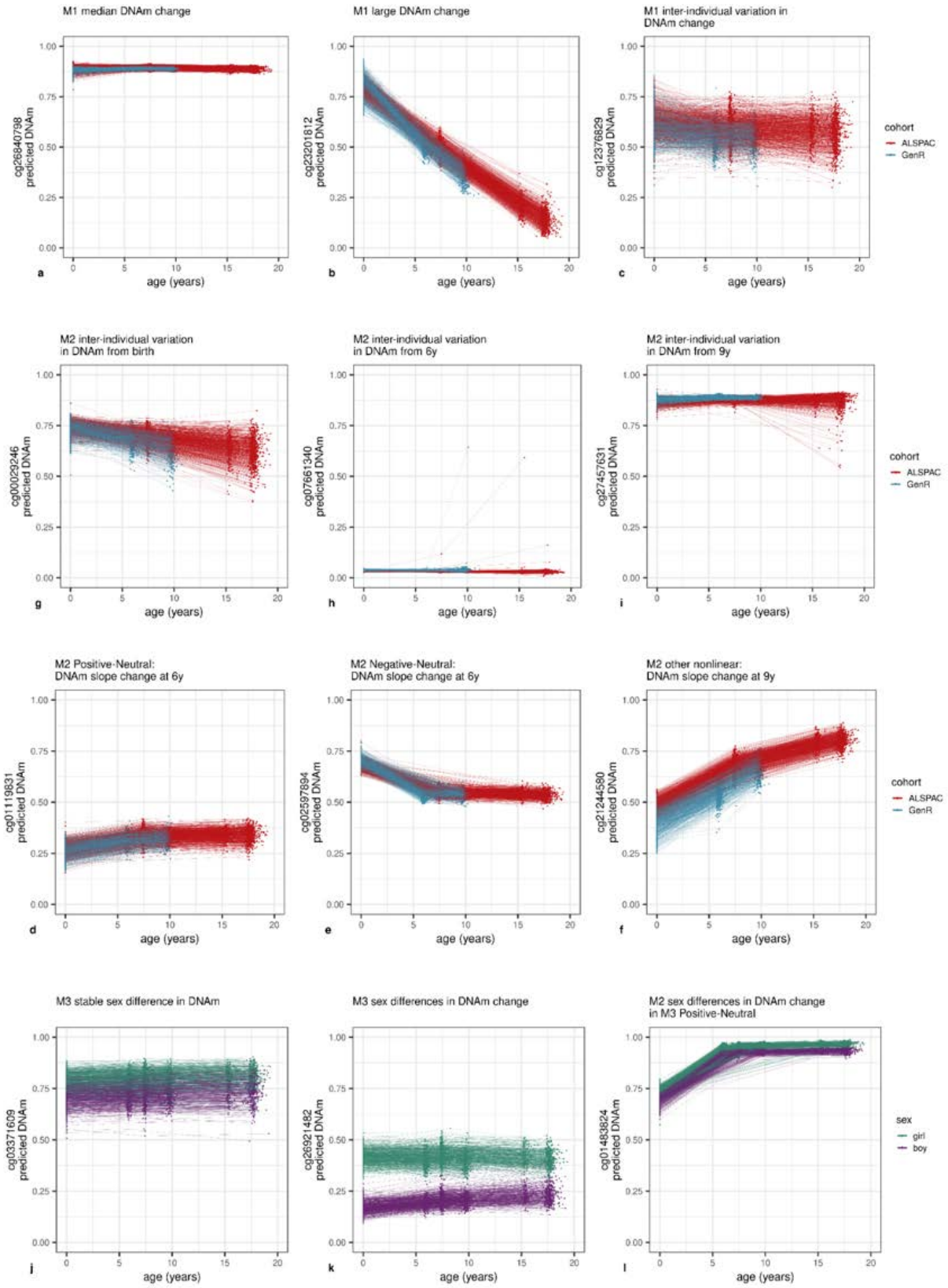


Figure 3. DNAm levels of selected CpG sites across childhood. Parts (a-c) show CpG sites with linear change over time (Model 1). A typical site is shown in (a), the site with the largest observed change in

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(b) and with inter-individual variation in DNAm change (c). Parts (d-f) show CpG sites with non-linear change (Model 2). A Positive-Neutral trajectory is shown in (d), a Negative-Neutral trajectory in (e) and a Positive-More Positive-Less Positive in (f). Parts (g-i) show CpG sites with inter-individual variation in change (Model 2). A site with slope variation from birth is shown in (g), slope change variation at 6 in (h) and slope change variation at 9 in (i). Parts (j-l) show CpG sites with sex-specific DNAm. A site with stable sex differences is shown in (Model 3) (j), sex-specific slope in (Model 3) (k) and sex-specific slope change at 6 in (Model 2) (l).

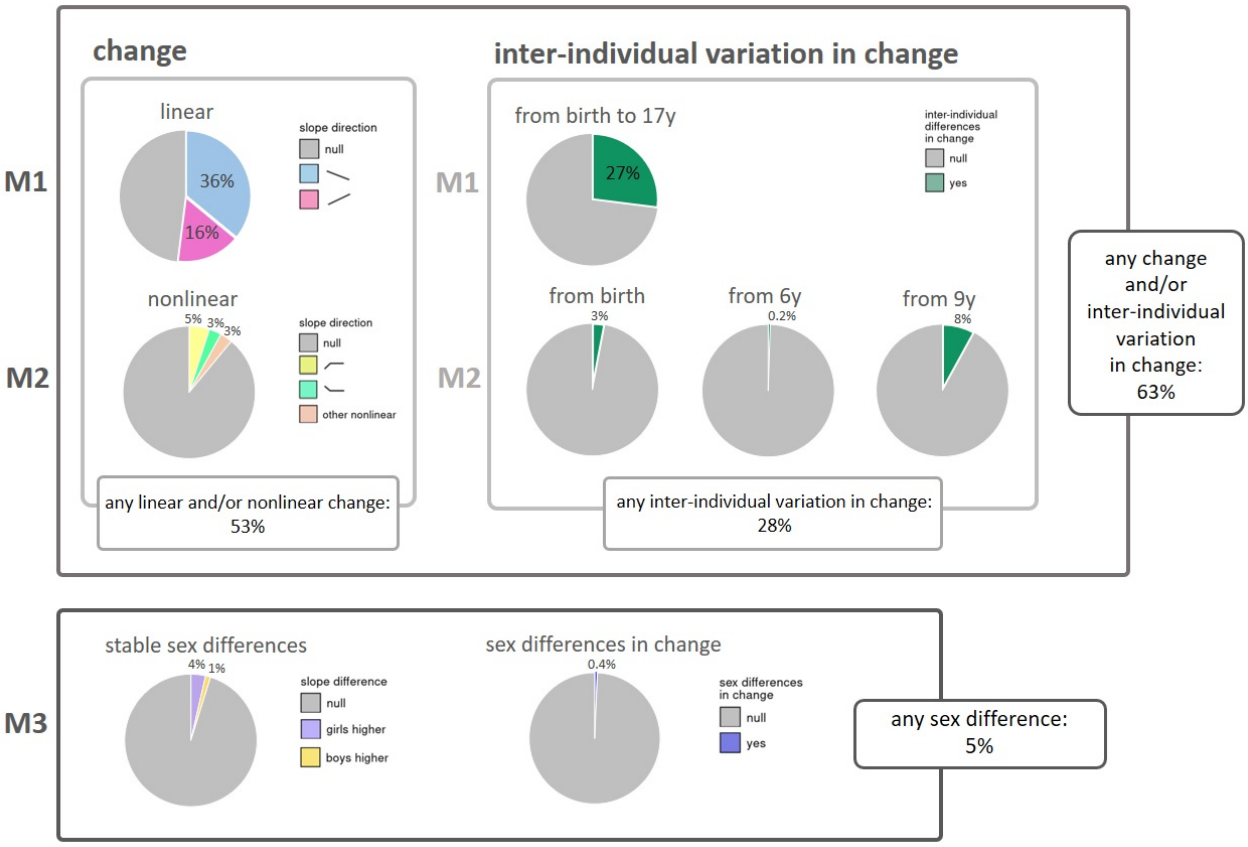


Figure 4. Overview of results from the three models. Model 1 (M1) was applied for overall change in DNA methylation and inter-individual variation in linear change; Model 2 (M2) for nonlinear change in DNA methylation and inter-individual variation in nonlinear change; and Model 3 (M3) for stable sex differences in DNA methylation and sex differences in change of DNA methylation (Sex by Time

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

interaction). Percentages represent percentage of autosomal CpGs below Bonferroni-corrected threshold ($P<1x10^{-07}$).

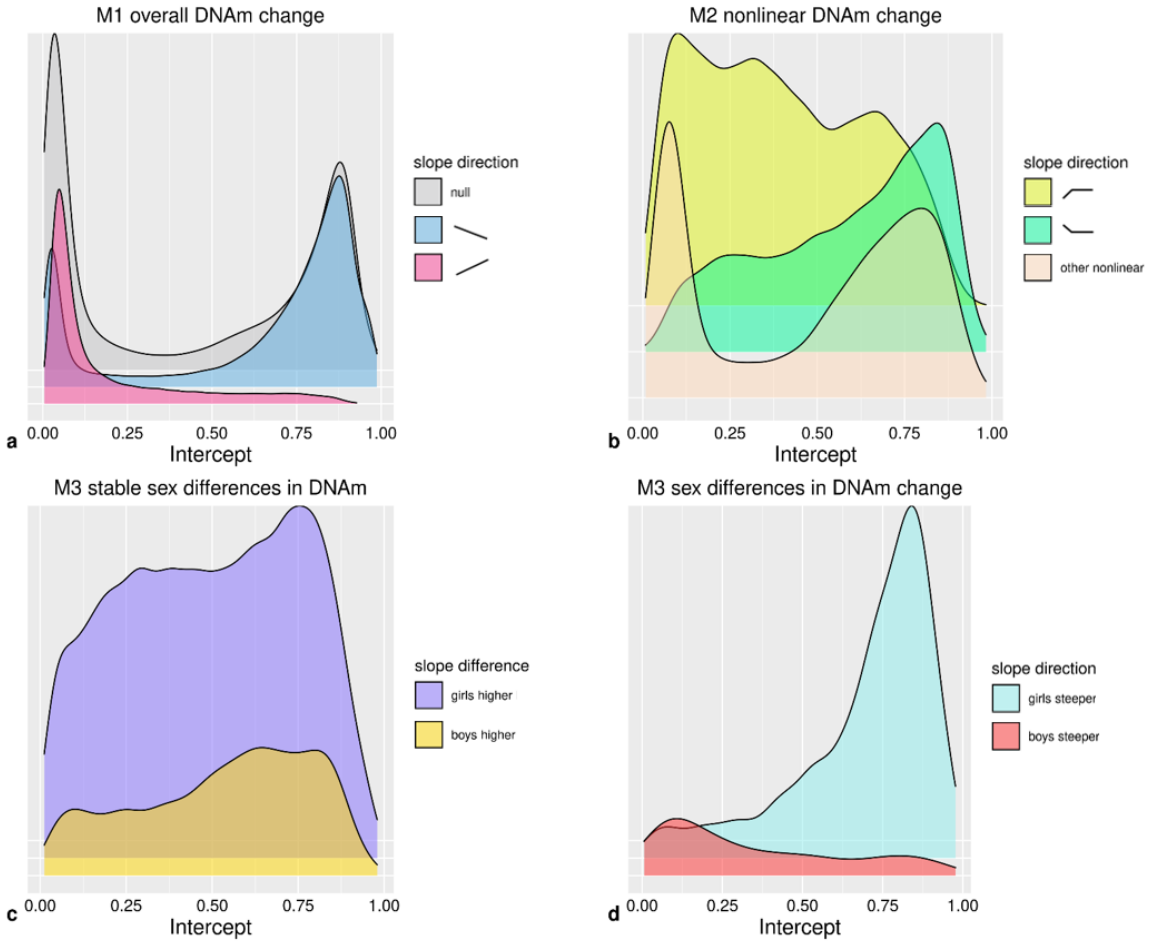


Figure 5. Density plots of intercepts of CpGs. Intercepts for CpGs with (a) directions of change in Model 1 (n=473,864); (b) nonlinear trajectories in Model 2 (n=52,043); (c) stable sex differences in Model 3 (n=22,821); (d) sex differences in DNAm change in Model 3 (n=1,768).

Nonlinear DNAm change

Model 2 (see online Methods) was identical to Model 1 but permitted slope changes at ages 6 and 9 years to test for nonlinear DNAm trajectories. At 11.0% of CpGs a nonlinear trajectory was detected. Specifically, at 4.8% of all CpGs, DNAm increased from birth and remained stable from 6 onward

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(Positive-Neutral; Figure 3d). Second, at 3.1% of all CpGs, DNAm decreased from birth and then remained stable at 6 years (Negative-Neutral; Figure 3e). The remaining 3.0% of all CpGs followed other nonlinear trajectories (e.g. Figure 3f), with each trajectory observed in <1.0% of all CpGs. Overall, linear and/or nonlinear changes in Model 1 or 2 were observed in 52.6% of CpGs (Figure 3), indicating that most nonlinear patterns were also detected as linear patterns in Model 1.

Inter-individual differences in change (i.e. random variance in slopes) from birth onward was detected at 3.4% of all sites (Figure 3g), inter-individual differences in slope change at 6 years in 0.2% (Figure 3h), and inter-individual differences in slope change at 9 years at 8.2% of CpGs (Figure 3i). Inter-individual differences in slope (change) at each time-point were detected more often at CpGs with an increasing rather than decreasing overall DNAm change in Model 1 (P=2.37x10^{-144}). Last, both Positive-Neutral and Negative-Neutral changes coincided more often with inter-individual variation from birth (P<9.88x10^{-324}). Any inter-individual differences in change, detected by Model 1 or 2, was observed at 27.9% of CpGs. In total, Models 1 and 2 detected age-related change whether linear, non-linear or inter-individual differences in change at 62.8% of all CpG sites (Figure 3).

Sex differences in longitudinal DNAm and DNAm change

According to Model 3 (online Methods), sex differences in DNAm were present at 4.9% of (autosomal) CpGs (Figure 3). Specifically, stable longitudinal sex differences (main sex effects) were observed at 4.8% of all (autosomal) CpGs (Figure 3j), and sex differences in DNAm change (sex by age interaction effects) were found at 0.4% of all (autosomal) CpGs (Figure 3k). At sites with stable sex differences, DNAm levels were higher in girls at 3.6% (Figure 3j) and lower at 1.2% of CpG sites. DNAm at sites with higher DNAm in girls tended to increase over time, whereas DNAm at sites with higher DNAm in boys tended to decrease (P=4.20x10^{-205}). Most commonly (at 0.2% of all CpGs), DNAm was higher in girls at birth but DNAm in boys increased at a higher rate.

Both CpGs with stable sex differences and those with sex differences in DNAm change were less likely to show inter-individual variation than other sites (20.8% versus 27.5% and 18.1% versus

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

27.3%; P=5.36x10^{-111} and P=7.57x10^{-18}). Finally, CpGs with stable sex differences or sex differences in DNAm change detected in Model 3 were much more likely to follow an overall Positive-Neutral trajectory of DNAm change detected in Model 2 than other CpG sites were (24.2% of CpGs with stable sex differences followed a Positive-Neutral trajectory versus 3.8% of other CpGs and 53.9% of CpGs with sex differences in DNAm change followed a Positive-Neutral trajectory versus 4.6% of other CpGs; P<9.88x10^{-324}, P<9.88x10^{-324}; Figure 3l). Albeit less prominently so, CpGs with stable sex differences or sex differences in DNAm change also more often followed a Negative-Neutral trajectory than other CpGs did (stable sex differences: 5.0% versus 3.0%, P=5.43x10^{-62}; sex differences in DNAm change: 7.7% versus 3.1%, P<7.11x10^{-28}).

Follow-up analyses

Follow-up analyses were performed to understand how different types of age-related DNAm trajectories are distributed across the genome (Supplementary Tables 3-5). All reported enrichments have significance below a Bonferroni-corrected threshold of P<4.46x10^{-04}, corrected for the number of chi-square tests (n=112). We further report enrichment of Gene Ontology (GO) pathways (nominal P<0.05) for genes annotated to CpG sites in each trajectory (Supplementary Tables 5-7). Last, we study enrichment of age-related DNAm trajectories in reported hits of different EWASs (Figure 6). All reported EWAS enrichments are below a Bonferroni-corrected threshold of P<2.16x10^{-04}, corrected for the number of Fishers' exact tests (n=231; Supplementary Table 8).

Patterns of DNAm change and CpG location

CpG sites with DNAm change associated patterns were labeled by gene associated regions, CpG island associated regions, as well as enhancer elements. Although many exceptions exist, low levels of DNAm in the promoter area but high levels of DNAm in the gene body are generally associated with increased gene transcription[32,33]. CpGs annotated to TSS200 regions more often showed an overall DNAm increase (Model 1) than other CpGs (19.0% versus 15.6%), whereas CpGs annotated to

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the gene body more often showed an overall DNAm decrease than other sites (38.8% versus 33.7%). TSS200 CpGs showed less inter-individual variation in overall DNAm change than other sites (22.2% versus 28.1%), whereas gene body CpGs showed somewhat more inter-individual variation in overall DNAm change than other sites (28.9% versus 26.5%).

Promoter areas often coincide with CpG islands[34]. Here, 63.3% of TSS200 CpGs were also annotated to CpG islands. As in TSS200 areas, CpGs annotated to CpG islands had lower DNAm levels (mode M1 intercept 2.4% (*SD*=30.2%)), and more often showed an overall DNAm increase than other sites (25.2% versus 12.0%). DNAm sex differences were especially present in the shores of CpG islands compared to all other island associated regions (stable sex differences: 7.5% versus 4.0%, sex differences in DNAm change: 0.6% versus 0.3%).

Enhancers act on promoters to regulate gene transcription[35]. CpGs annotated to enhancer elements (2.0% of CpGs) tended to have low DNAm levels (mode M1 intercept 5.07%; *SD*=31.4%) and then increased with age more than other CpGs (23.9% versus 15.9%). Inter-individual variation in change from birth was more common at enhancer sites than at other sites (5.6% versus 3.3%).

Functional associations

Enrichment of Gene Ontology categories was tested for genes linked to CpGs with different DNAm trajectories. In short, genes annotated to CpGs with overall decreasing DNAm levels were enriched in immune-developmental functions, whereas those annotated to CpGs with increasing levels were enriched in neurodevelopmental functions. This pattern seemed even more pronounced at genes annotated to nonlinear Negative-Neutral and Positive-Neutral CpGs, with the former more often associated to immune-development and the latter to neurodevelopment. Genes linked to CpGs with stable sex differences and sex differences in DNAm change were enriched in pathways associated with sexual development, such as genital development, as well as pathways associated with neurodevelopment. Genes linked to CpGs with sex difference in DNAm change were also enriched in functions related to tooth and hair development.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Enrichment in EWASs

We further investigated functional relevance of CpG sites with age-related DNAm trajectories by testing enrichment with published EWAS associations (Figure 6)[28,36-61]. Unsupervised clustering of the enrichments shows that CpG sites with inter-individual variation in change over time have distinct enrichments and cluster differently from those with age-associated change that is consistent among individuals. The CpG sites of each age-associated DNAm trajectory were enriched with published age associations in adulthood. Multiple smoking EWAS clustered together with enrichment patterns exhibiting strongest enrichments among CpG sites with Negative-Neutral trajectories and mostly weak enrichments among CpG sites with inter-individual variation in change. Further, despite adjusting for cell count heterogeneity in our models, we observed enrichments of CpG sites that differ by white blood cell type among sites following nearly all age-associated trajectories. Finally, we observed enrichments of CpG sites associated with gestational age and prenatal smoking with sex-specific DNAm.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

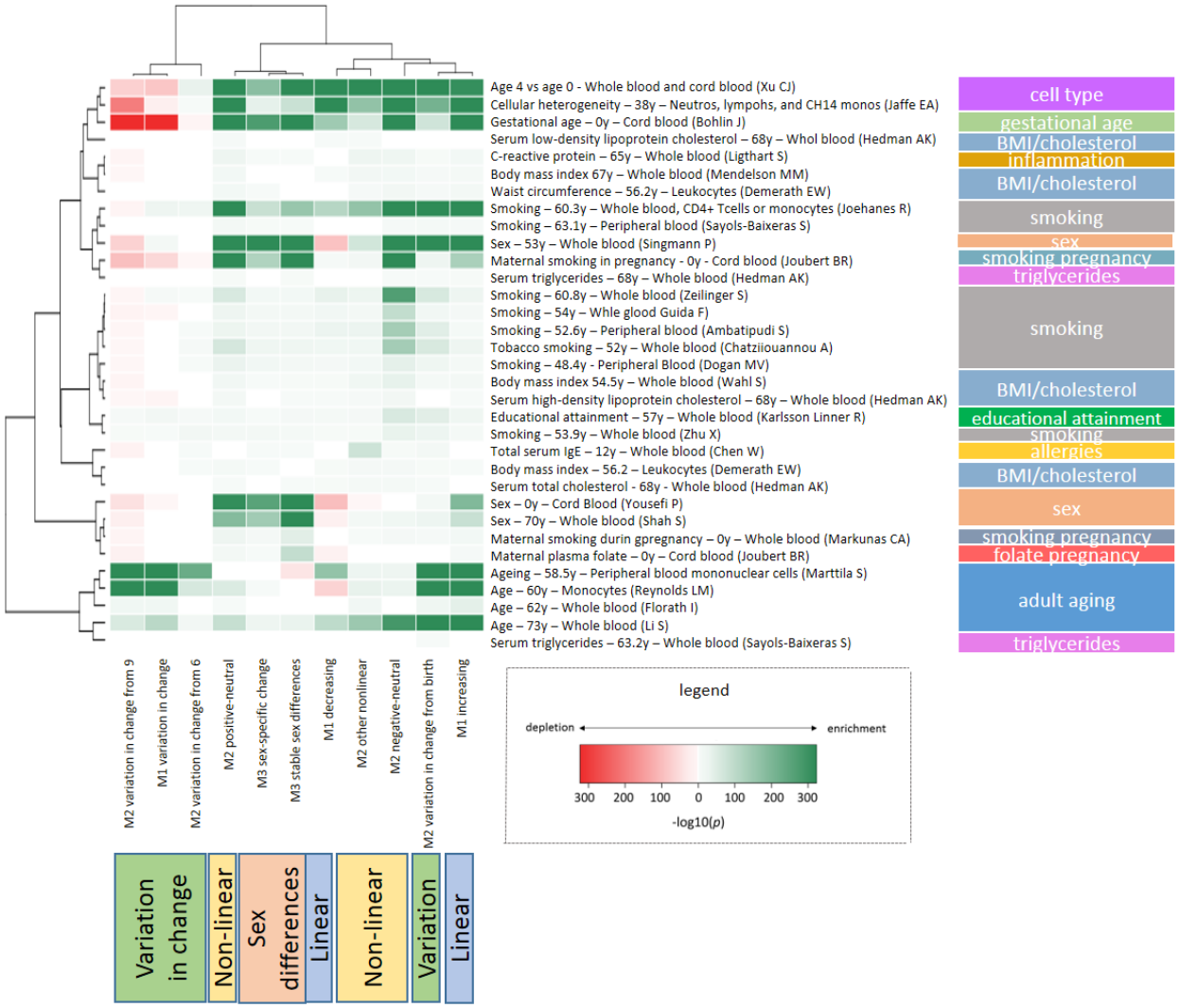


Figure 6. Enrichment of age-related trajectories in EWASs

Discussion

In this study we described changes in DNAm levels through the first two decades of human life. We examined DNAm levels per CpG by their linear association with age, their nonlinear trajectories and inter-individual variation in change, as well as sex differences and CpG characteristics.

We found that about half of sites change: consistent linear and/or nonlinear DNAm change was found at 53% of sites. We further found that over a quarter of sites, 28%, were characterized by substantial inter-individual differences in the direction of this change. DNAm sex differences were

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

present, but not abundant: 5% of autosomal sites displayed different DNAm levels or differences in change over time for girls and boys.

Specifically, we determined that DNAm at 52% of the measured methylome have some form of linear change from birth to late adolescence, with DNAm decreasing at 36% and increasing at 16% of CpGs. CpGs with decreasing DNAm tended to have high levels of DNAm and were more often located in gene bodies. CpGs with increasing levels of DNAm tended to have low levels of DNAm and were more likely to be located in promoter regions and at enhancers. The predominance of decreasing CpGs is in agreement with literature on epigenome-wide DNAm and age in cross-sectional research on children and adults[18,62], as well as with longitudinal research in adults[19].

Nonlinear DNAm trajectories were detected at 11% of CpGs, mostly involving changes in DNAm from birth to age 6 years, after which DNAm was more stable. We note that this could be due to cord blood being used to generate DNAm profiles at birth, whereas peripheral blood was used at later ages. A previous study[23] including eight children showed that the cord blood DNAm profile at birth clustered separately from later peripheral profiles, after which DNAm changed gradually from 1, to 2.5, to 5 years. Such differences between DNAm in cord and peripheral blood might be due to uncaptured differences in white blood cell composition, as well as to different gene-regulatory functioning in the intra-uterine versus extra-uterine environment. On the other hand, Gene Ontology analyses showed that functional associations for positive and negative *linear* DNAm patterns, which are unlikely to be affected by tissue type, were consistent with functional associations for nonlinear positive and negative patterns, respectively (e.g. positive and negative up to 6 years, and then no change up to 18 years).

Specifically, sites with decreasing levels of DNAm, both with or without slope changes around the age of 6 years, were functionally enriched for immune-developmental pathways, and sites with increasing levels of DNAm, both with or without slope changes, were enriched for neurodevelopmental pathways. Since these observations were based on blood DNAm, it remains to

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

be studied what roles genes linked to neurodevelopmental pathways play in in blood, or to what extent DNAm trajectories in blood mirror those in neural tissue.

Inter-individual differences in linear DNAm trajectories were found at 27% of CpGs, indicating change at different rates or directions for different individuals. Such sites tended to have overall increasing rather than decreasing levels of DNAm from birth to 18 years. This observation is consistent with the only other large study to examine inter-individual differences in DNAm change[20], although we note that this study included elderly subjects and detected a smaller number of relevant CpGs. We are the first to investigate inter-individual differences in nonlinear DNAm trajectories. These were most often found in the slope change at 9 years (8% of CpG sites), indicating that most inter-individual differences in DNAm emerge after the first decade of life. More research is needed to understand if the direction of change in this period is determined by stimuli during that period, or rather by preceding, perhaps cumulative, exposures. However, it is clear that, given the high proportion of CpG sites with inter-individual variation in DNAm change over time that we have observed, it is important to restrict the range of ages of children included a single EWAS. Specific limits should be discussed given the rapidly growing number of studies generating DNAm profiles across childhood[63].

Stable sex differences were found at 5% of autosomal CpGs, and sex differences in DNAm *change* were found at 0.4% of all CpGs. In general, if there were stable sex differences, girls had higher levels of DNAm (4% of all CpGs), in case of sex differences in DNAm change, boys had an accelerated upward change (0.2% of all CpGs). The direction of stable sex differences detected are congruent with a cross-sectional study on newborns, in which girls had higher DNAm levels than boys for the large majority of the 3031 significant autosomal CpGs[54]. Sex-discordant associations with age seemed to be more prevalent from birth to age 6 years than afterwards, suggesting that any phenotypic sex differences associated to DNAm would be established in early childhood. Their enrichment in the shores of CpG islands, areas at which DNAm has been associated with tissue differentiation and tissue-specific gene expression[64], is consistent with the critical role that these

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

processes play in sexual differentiation. Studies into sex differences in epigenetic regulation might want to focus on these locations.

We also found the other DNAm trajectories to be arranged throughout the genome in a nonrandom fashion. Earlier studies[32,65] have shown that, for active genes, lower DNAm towards the promoter area (TSS200) and higher DNAm in the gene body relate to increased gene transcription. Here we add the observation that promoter DNAm tends to increase and gene body DNAm tends to decrease with age. From this finding, one might infer that a downregulation of gene expression takes place from birth to late adolescence. Enrichment analyses of published EWAS associations further showed that different traits and exposures exhibited distinct enrichment patterns among DNAm trajectories. For example, there were clear differences between smoking and BMI-related traits. Enrichment of sites with DNAm sex differences in EWASs on prenatal maternal smoking is consistent with studies finding that prenatal smoking affects traits such as birth weight[66] , brain development[67,68], and attention[69] differently in boys and girls. Clustering for prenatal maternal smoking EWASs also showed enrichment for CpGs with consistent change among individuals, not for CpGs with inter-individual variation in change. This may suggest a link with the well-known effects of prenatal smoking on childhood development since consistent DNAm change is more likely related to development or aging programming than inter-individual variation. This may explain why changes associated to prenatal smoking persist throughout life[70]. Notably, this pattern of change without inter-individual variation is visible in cg05575921, the *AHRR* CpG site strongly and persistently associated with prenatal smoking[71,72] (Supplemental Figure 2; http://epidelta.mrcieu.ac.uk/).

'Epigenetic age acceleration' is a term coined to indicate the deviation of chronological age from age as estimated by an 'epigenetic clock' and is associated with disease risk and mortality[73]. Existing clocks are all linear models based on DNAm. Consequently, one might expect that all CpGs included in the clock model change linearly with age. Furthermore, to detect age acceleration, one would expect that these CpG sites would also vary between individuals. Surprisingly, many CpG sites included in the most popular clocks do not match these expectations[74,75] (Supplemental Tables 9, 10).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

For example, we observe that over one-quarter and nearly one-half of the CpG sites included in the Horvath and Hannum clocks, respectively, follow non-linear DNAm trajectories in childhood. Given the widespread use of clocks to investigate biological aging, further investigation is warranted to better understand how, and perhaps if, associations using these clocks should be interpreted in child DNAm profiles.

We note three main limitations of our findings. First, the use of different tissue types (cord blood and peripheral blood) could account for some of the differences between birth and later time points, e.g. sites that increased or decreased between birth and 6, but did not show change after that. Generation of DNAm profiles of a single tissue or cell type collected across childhood would be needed to disentangle this issue further. Unfortunately, such a dataset is not currently available as most cohorts have generated DNAm profiles from peripheral blood and cord blood[63]. Analysis of these complex tissues has nevertheless yielded many valuable insights. Second, since DNAm at 9 years was measured only in Generation R and at 17 years only in ALSPAC, DNAm differences from 9 to 17 may be to some extent driven by batch effects or cohort differences. This may explain some of the inter-individual differences in slope changes at 9 towards 17 years. However, the high level of agreement in both stability and change among the corresponding time points of the two cohorts is reassuring. Moreover, it is not entirely surprising that inter-individual variation in directionality of change was higher for the largest age interval. This interval, furthermore, encompasses the period of adolescent development, a time in which many inter-individual phenotypic differences arise. Finally, it should be noted that the current study only included children of European ancestry. Considerable DNAm differences have been found between populations[76-78], but research on age-associated DNAm differences is scarce. One study[79] reported evidence for overlap in age-associated CpGs in two African populations with studies on European-ancestry populations, but more research is needed to map the generalizability of longitudinal DNAm changes among different populations.

In conclusion, in the first comprehensive CpG-by-CpG characterization of DNAm from birth to late adolescence, we found that DNAm at more than half of the studied CpG sites changes

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

consistently between individuals, and that considerable inter-individual variation in change exists. Further, characteristics such as child sex, CpG location, and environmental and disease traits have distinct associations with patterns of DNAm change. Further analysis of these patterns is made readily available at http://epidelta.mrcieu.ac.uk/, which we hope can be used in future studies to test developmental hypotheses that promote our understanding of the developmental nature of DNAm, its role in gene functioning, and the associated biological pathways leading to health and disease.

Methods

Setting

Data were obtained from two population-based prospective birth cohorts, the Dutch Generation R Study (Generation R) and the British Avon Longitudinal Study of Parents and Children (ALSPAC). Pregnant women residing in the study area of Rotterdam, the Netherlands, with an expected delivery date between April 2002 and January 2006 were invited to enroll in Generation R. A more extensive description of the study can be found elsewhere[80]. The Generation R Study is conducted in accordance with the World Medical Association Declaration of Helsinki and has been approved by the Medical Ethics Committee of the Erasmus Medical Center, Rotterdam. Informed consent was obtained for all participants.

Pregnant women residing in the study area of former county Avon, United Kingdom, with an expected delivery date between April 1991 and December 1992 were invited to enroll in the ALSPAC study. Detailed information on the study design can be found elsewhere[3,81]. The ALSPAC website contains details of all available data through a fully searchable data dictionary and variable search tool (http://www.bristol.ac.uk/alspac/researchers/our-data/). Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Informed consent for the use of data collected via questionnaires and clinics was obtained from participants following the recommendations of the ALSPAC Ethics and Law Committee at the time.

Study Population

In the Generation R Study, 9,778 pregnant mothers had 9,749 live-born children. For a subsample of 1,414 children DNAm data was collected at birth and/or 6 years and/or 10 years of age. This subsample consisted of participants with parents born in the Netherlands (European ancestry[82] confirmed for all children with genetic data available (95.4%)). Fifteen sibling pairs were present in the dataset. From each pair one sibling with the lowest number of DNAm measurements, or otherwise randomly, was excluded, resulting in a sample with 1,399 children (with 2,333 DNAm samples; see below).

In the ALSPAC study, 15,247 pregnant mothers gave birth to 14,973 live-born children. DNAm at birth and/or 7 years and/or 17 years was available for a subsample of 1,003 children as part of the Accessible Resource for Integrated Epigenomic Studies (ARIES) study[83]. From this sample, 48 children with non-European ancestry as based on genetic principle component analysis and 6 children with missing data on gestational age were excluded, resulting in a sample of 949 children with DNAm data (with 2,686 DNAm samples; see below).

DNA methylation

Cord blood was drawn after birth for both cohorts, and peripheral blood was drawn at a mean age of 6.0 (*SD*=0.47) and 9.8 (*SD*=0.3) years for Generation R, and 7.5 (*SD*=0.2) and 17.1 (*SD*=1.0) years for ALSPAC. Both cohorts made use of the EZ-96 DNAm kit (shallow) (Zymo Research Corporation, Irvine, USA) to perform bisulfite conversion on the extracted leukocytic DNA. Samples were further

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA) to analyze DNAm.

In Generation R, quality control was performed on all 2,467 available DNAm samples with the CPACOR workflow[84]. Arrays with observed technical problems such as failed bisulfite conversion, hybridization or extension, as well as arrays with a mismatch between sex of the proband and sex determined by the chromosome X and Y probe intensities were removed from subsequent analyses. Additionally, only arrays with a call rate >95% per sample were processed further, resulting in 2,355 samples, 22 of which belonged to half of an excluded sibling pair, hence 2,333 samples were carried forward into normalization.

In ALSPAC, quality control was performed on 6,057 samples (3,286 belonging to children, 2,771 to their mothers), using the *meffil* package[85] in R version 3.4.3[86]. After removing samples with mismatched genotypes, mismatched sex, incorrect relatedness, low concordance with samples collected at other time points, extreme dye bias, and poor probe detection, 5,337 samples remained, 2,845 of which belonging to children, used in the current study.

To minimize cohort effects as much as possible, we normalized both cohorts together as a single dataset. Functional normalization (10 control probe principal components, slide included as a random effect) was performed with the *meffil* package in R[85]. Normalization took place on the combined Generation R and ALSPAC set comprising a total of 5,178 samples for a total of 485,512 CpGs. One-hundred and fifty-nine ALSPAC samples belonging to non-European children or children with missing data on gestational age were excluded, leading to a final ALSPAC set of 2,686 samples (for 949 children). Together with 2,333 samples for Generation R (of 1,399 children) they formed a combined set of 5,019 samples (of 2,348 children.)

Analyses were restricted to 473,864 autosomal CpGs. DNAm levels were operationalized as beta values (β values), representing the ratio of methylated signal relative to the sum of methylated and unmethylated signal measured per CpG.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Covariates

Sample plate number (*N*=29 in Generation R and *N*=31 in ALSPAC), was used to correct for batch effects, which was added as a random variable in the model (see below). White blood cell (WBC) composition was estimated with the reference-based Bakulski method[87] for cord blood and Houseman method[88] for peripheral blood (Supplemental Table 11). Nucleated red blood cells were not further analyzed due to its specificity to cord blood, leaving CD4+ T-lymphocytes, CD8+ T-lymphocytes, natural killer cells, B-lymphocytes, monocytes, and granulocytes. Other covariates included gestational age in weeks, sex of the child, and cohort.

Statistical analyses

Step 1: Assessing cross-cohort comparability in DNA methylation stability

To ascertain comparability amongst the two cohorts we compared within-cohort DNAm stability between the time points that were present in both cohorts – i.e. birth and 6/7 years (Generation R/ALSPAC, respectively).

Longitudinal stability per CpG within each cohort was assessed by studying estimates of concordance and change. For concordance, DNAm data was first residualized within each cohort for all variables present in the longitudinal models except the 'cohort' variable, in order to remove between-cohort differences due to other covariates. Concordance was then measured both with Spearman correlation (data at most CpGs is not normally distributed) as a measure of relative concordance, and with intra-class correlations as a measure of absolute concordance (children with data for both time points: *n* Generation R=476, *n* ALSPAC=826). Longitudinal change from birth to 6/7 years was assessed by studying the estimates of the change in DNAm per year by applying Model 1 (see below) within each cohort (children with data for at least one of the two time-points: *n* Generation R=1,394, *n* ALSPAC=944).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

In a second step, cross-cohort comparability was assessed with Spearman (ρ) correlation of concordance estimates of the CpGs of each cohort (which were not normally distributed) and Pearson correlations (*r*) amongst the change estimates of the CpGs of each cohort (which were normally distributed).

Step 2: Longitudinal modelling of DNA methylation using combined Generation R and ALSPAC data

The combined Generation R and ALSPAC dataset had four time points of collection (birth, age 6/ 7 years, 10 years, and 17 years). We fit three linear mixed models to CpG site DNAm across the genome to identify (i) linear change over time (Model 1); (ii) nonlinear change over time (Model 2); and (iii) sex differences in change over time (Model 3). Both fixed and random effects were examined to allow for inter-individual variation in DNAm patterns over time. The models are described in detail below.

Model 1: Linear change. This model was applied to identify CpGs that show an overall change in DNAm from birth to 18 years (i.e. fixed age effect), as well as CpGs with inter-individual differences in change during that time (i.e. random age effect). The Model 1 is defined as:

M1: $M_{ijk} = \beta_0 + u_{0i} + \beta_1 Age_{ij} + u_{1i} Age_{ij} + u_{0k} + \text{covariates} + \epsilon_{ijk}$

$\epsilon_{ijk} \sim N(0, \sigma_\epsilon^2)$

$u_{0i} \sim N(0, \sigma_{0i}^2)$

$u_{1i} \sim N(0, \sigma_{1i}^2)$

$u_{0k} \sim N(0, \sigma_{0k}^2)$

Here, participants are denoted by *i*, time points by *j,* and sample plates by *k*. *M* denotes DNAm level, β_0 fixed intercept, u_{0i} random intercept, β_1 fixed age coefficient, u_{1i} random age coefficient, u_{0k} random intercept for sample plate. Hence, β_1 represents the average change in DNAm per one year.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Variability in this change amongst individuals was captured with u_{1i}. To avoid problems with model identification, the random slope of age was uncorrelated to the random intercept (i.e. a diagonal random effects matrix was used).

Model 2: Nonlinear change. To identify nonlinear changes in DNAm, we extended Model 1 to allow slope changes at ages 6 and 9[30,31]:

M2: $M_{ijk} = \beta_0 + u_{0i} + \beta_1 Age_{ij} + \beta_2 (Age_{ij}-6)^+ + \beta_3 (Age_{ij}-9)^+ + u_{1i} Age_{ij} + u_{2i} (Age_{ij}-6)^+ + u_{3i} (Age_{ij}-9)^+ + u_{0k} + \text{covariates} + \epsilon_{ijk}$

$\epsilon_{ijk} \sim N(0,\sigma_\epsilon^2)$

$u_{0i} \sim N(0,\sigma_{0i}^2)$

$u_{1i} \sim N(0,\sigma_{1i}^2)$

$u_{2i} \sim N(0,\sigma_{2i}^2)$

$u_{3i} \sim N(0,\sigma_{3i}^2)$

$u_{0k} \sim N(0,\sigma_{0k}^2)$

Where $a^+ = a$ if $a>0$ and 0 otherwise, so that β_2 represents the average change in DNAm per year from 6 years of age onward, after accounting for the change per year from birth onward, as denoted by β_1. Likewise, $\beta_3$represents the average change in DNAm per year from 9 years of age onward, after accounting for the change per year from 6 years of age onward. Hence, with those variables we are able to detect slope changes at 6 and 9 years old. These slope changes were used to identify different types of nonlinear patterns. With u_{2i} and u_{3i} the inter-individual variation in slope changes at 6 and 9 years were captured, respectively. General linear hypothesis testing[89] was applied to our fitted models to determine if there were changes in DNAm per year from 6-9 years and from 9-18 years.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Model 3: Sex differences in change: To identify CpGs for which DNAm changes differently over time for boys and girls, we applied the following model:

M3: $M_{ijk} = \beta_0 + u_{0i} + \beta_1 Age_{ij} + u_{1i} Age_{ij} + \beta_2 Sex_i Age_{ij} + u_{0k} + \text{covariates} + \epsilon_{ijk}$

$\epsilon_{ijk} \sim N(0,\sigma_\epsilon^2)$

$u_{0i} \sim N(0,\sigma_u^2)$

$u_{1i} \sim N(0,\sigma_{1i}^2)$

$u_{0k} \sim N(0,\sigma_{0k}^2)$

Here, Sex_i denotes the sex of child *i*. Both main and interaction effects for sex were studied.

The three mixed models were fitted using maximum likelihood estimation in *R* with the *lme4* package[90]. Continuous covariates (WBCs, gestational age) were z-score standardized. Random slopes were kept uncorrelated with random intercepts and the NLopt optimizer was used, enabling us to improve computational speed compared to the default settings. *P*-values for the fixed effects were computed with a z-test. *P*-values for random slopes of the Age effects were obtained by refitting the model without the random slope and comparing the fit estimates of the two models with a likelihood ratio test. Within each model, *P*-value thresholds were Bonferroni-corrected for the number of tested CpGs (i.e. to $P<1x10^{-07}$).

Step 3: Functional characterization of probes with comparable patterns of change

To interpret the functionality of the age-related DNAm patterns from the three models, CpG sites adhering to 8 different age-related patterns (M1 linear change and inter-individual variation in linear change, M2 nonlinear trajectories, and inter-individual variation in change from birth, in slope change at 6 years, and in slope change at 9 years, and M3 stable sex differences and sex differences in DNAm change) were tested for enrichment in:

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(i) gene-relative genomic regions (TSS1500, TSS200, 5'UTR, 1st exon, gene body, 3'UTR, and intergenic regions[64]),

(ii) CpG island-relative genomic regions (N shelf, N shore, CpG island, S shore, S shelf, and open sea regions[64]) as indicated by the Illumina HumanMethylation450 v1.2 Manifest (Illumina Inc., San Diego, USA), and

(iii) enhancer elements as those expressed in whole blood, peripheral blood mononuclear cells, natural killer cells, CD4+ T cells, CD8+ T cells, monocytes, neutrophils, eosinophils or B cells[91],

Altogether, these encompassed 14 enrichment analyses for 8 variables. Enrichment was tested using χ^2-tests of unequal proportions. The enrichment *P*-value threshold was Bonferroni-corrected for multiple tests (i.e. $P<4.46\times10^{-04}$ for 8x14=122 tests). Second, we tested enrichment of Gene Ontology (GO) categories for genes linked to CpG sites surviving adjustment for multiple tests ($P<1\times10^{-07}$) for each of the main variables of interest. This analysis was adjusted for gene size and pruned for near-identical terms (see elsewhere for a full description[92]). For completeness, terms with nominal $P<0.05$ were reported. Last, we tested enrichment of age-related DNAm trajectories (11 different age-related patterns: M1 decreasing, increasing, and inter-individual variation in linear change, M2 Positive-Neutral, Negative-Neutral, other nonlinear, inter-individual variation in change from birth, in slope change at 6 years, and in slope change at 9 years, and M3 stable sex differences and sex differences in DNAm change) in EWASs on age, prenatal smoking, smoking, cardiovascular-associated traits, C-reactive protein, allergies, educational attainment, and cellular heterogeneity. EWAS summary statistics were retrieved from the EWAS Catalog (http://www.ewascatalog.org/) and studies were included when performed with the 450K array in peripheral or cord blood, resulting in 21 EWASs. Enrichment was tested with Fisher's exact tests, the enrichment *P*-value threshold was Bonferroni-corrected for multiple tests (i.e. $P<1.37\times10^{-04}$ for 11x33=363 tests).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Data access

Data from the Generation R Study are available upon request (generationr@erasmusmc.nl), subject to local rules and regulations. ALSPAC data access is through a system of managed open access. The ALSPAC access policy (http://www.bristol. ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf) describes the process of accessing the data and samples in detail, and outlines the costs associated with doing so. The results per CpG will be made available at http://epidelta.mrcieu.ac.uk.

Acknowledgements

The Generation R Study is conducted by the Erasmus Medical Center in close collaboration with the Faculty of Social Sciences of the Erasmus University Rotterdam, the Municipal Health Service Rotterdam area, Rotterdam, the Rotterdam Homecare Foundation, Rotterdam and the Stichting Trombosedienst & Artsenlaboratorium Rijnmond (STAR-MDC), Rotterdam. We gratefully acknowledge the contribution of children and parents, general practitioners, hospitals, midwives and pharmacies in Rotterdam. The study protocol was approved by the Medical Ethical Committee of the Erasmus Medical Centre, Rotterdam. Written informed consent was obtained for all participants. The generation and management of the Illumina 450K methylation array data (EWAS data) for the Generation R Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Mr Michael Verbiest, Ms Mila Jhamai, Ms Sarah Higgins, Mr Marijn Verkerk and Dr. Lisette Stolk for their help in creating the EWAS database. We thank Dr A. Teumer for his work on the quality control and normalization scripts. The general design of the Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam, the Netherlands Organization for Health Research and Development and the Ministry of Health, Welfare and Sport. The EWAS data was funded by a grant from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

810), by funds from the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, and by a grant from the National Institute of Child and Human Development (R01HD068437). This project further received funding from the European Union's Horizon 2020 research and innovation programme (733206, LifeCycle; 848158, EarlyCause; 874739, LongITools) and from the European Joint Programming Initiative "A Healthy Diet for a Healthy Life" (JPI HDHL, NutriPROGRAM project, ZonMw the Netherlands no.529051022 and PREcisE project ZonMw the Netherlands no.529051023).

We are extremely grateful to all the families who took part in the ALSPAC study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. The UK Medical Research Council and Wellcome (grant number 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This publication is the work of the authors and Dr Suderman will serve as guarantor for the ALSPAC-related contents of this paper. Analysis of the ALSPAC data was funded by UK Economic & Social Research Council grant (grant number ES/N000498/1). ARIES was funded by the BBSRC (BBI025751/1 and BB/I025263/1). Supplementary funding to generate DNA methylation data which are (or will be) included in ARIES has been obtained from the MRC, ESRC, NIH and other sources. ARIES is maintained under the auspices of the MRC Integrative Epidemiology Unit at the University of Bristol (grant numbers MC_UU_00011/4 and MC_UU_00011/5).

AN and HT are supported by a grant of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant No. 024.001.003, Consortium on Individual Development). HT was also supported by a grant from the Netherlands Organization for Scientific Research (NWO-grant 016.VICI.170.200). EW is funded by CLOSER, who is funded by ESRC and MRC (grant reference: ES/K000357/1). The funders took no role in the design, execution, analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or interpretation of the data or in the writing up of the findings. www.closer.ac.uk. The work of CAMC has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 707404. LCH is supported by ESRC/BBSRC Grant ES/N000382/1 for the Interpreting epigenetic signatures in studies of early life adversity project (Interstela Project). JR is supported by the Netherlands Organization for Scientific Research (NWO ZonMw VENI, grant no 91618147). VWVJ received funding from the European Research Council (ERC-2014-CoG-648916). MJB-K is supported by the European Research Council (ERC AdG) and by the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant No. 024.001.003, Consortium on Individual Development). MHvIJ is supported by the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant No. 024.001.003, Consortium on Individual Development) and by a Spinoza Award of the Netherlands Organization for Scientific Research.

Author contributions

RHM, AN, EW, CAMC, LCH, JR, BTH, MJB-K, HT, MvIJ, CLR, and MS designed the study; RHM, AN, JR, TRG, JFF, HT, VWJ and CLR acquired the data; RHM and MS analyzed the data; RHM, AN, EW, CAMC, LCH, AJS, BTH, JFF, MJB-K, MvIJ, and MS interpreted the data; RHM drafted the work and all authors substantively revised it. All authors have approved the submitted version. All authors have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and the resolution documented in the literature. The funders took no role in the design, execution, analysis or interpretation of the data or in the writing up of the findings.

Disclosure declaration

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1 Khavari, D. A., Sen, G. L. & Rinn, J. L. DNA methylation and epigenetic control of cellular differentiation. *Cell cycle* **9**, 3880-3883 (2010).

2 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770 (2008).

3 Fraser, A. *et al.* Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *International Journal of Epidemiology* **42**, 97-110 (2012).

4 Chow, J. & Heard, E. X inactivation and the complexities of silencing a sex chromosome. *Current opinion in cell biology* **21**, 359-366 (2009).

5 Koukoura, O., Sifakis, S. & Spandidos, D. A. DNA methylation in the human placenta and fetal growth. *Molecular medicine reports* **5**, 883-889 (2012).

6 Eden, A., Gaudet, F., Waghmare, A. & Jaenisch, R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science* **300**, 455-455 (2003).

7 Comet, I., Riising, E. M., Leblanc, B. & Helin, K. Maintaining cell identity: PRC2-mediated regulation of transcription and cancer. *Nature Reviews Cancer* **16**, 803 (2016).

8 Natoli, G. Maintaining cell identity through global control of genomic organization. *Immunity* **33**, 12-24 (2010).

9 Bojesen, S. E., Timpson, N., Relton, C., Smith, G. D. & Nordestgaard, B. G. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax* **72**, 646-653 (2017).

10 Cao-Lei, L. *et al.* DNA methylation signatures triggered by prenatal maternal stress exposure to a natural disaster: Project ice storm. *PLoS ONE* **9**, doi:10.1371/journal.pone.0107653 (2014).

11 Baccarelli, A. *et al.* Rapid DNA methylation changes after exposure to traffic particles. *American Journal of Respiratory and Critical Care Medicine* **179**, 572-578 (2009).

12 Suarez-Alvarez, B., Rodriguez, R. M., Fraga, M. F. & López-Larrea, C. DNA methylation: a promising landscape for immune system-related diseases. *Trends in Genetics* **28**, 506-514 (2012).

13 Szulwach, K. E. *et al.* 5-hmC–mediated epigenetic dynamics during postnatal neurodevelopment and aging. *Nature Neuroscience* **14**, 1607 (2011).

14 Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biology* **17**, 61 (2016).

15 Feil, R. & Fraga, M. F. Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics* **13**, 97 (2012).

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

17 Hernandez, D. G. *et al.* Distinct DNA methylation changes highly correlated with chronological age in the human brain. *Human molecular genetics* **20**, 1164-1172 (2011).

18 Alisch, R. S. *et al.* Age-associated DNA methylation in pediatric populations. *Genome research* **22**, 623-632 (2012).

19 Wang, Y. *et al.* Epigenetic influences on aging: a longitudinal genome-wide methylation study in old Swedish twins. *Epigenetics* **13**, 975-987 (2018).

20 Wang, Y., Pedersen, N. L. & Hägg, S. Implementing a method for studying longitudinal DNA methylation variability in association with age. *Epigenetics* **13**, 866-874 (2018).

21 Lévesque, M. L. *et al.* Genome-wide DNA methylation variability in adolescent monozygotic twins followed since birth. *Epigenetics* **9**, 1410-1422 (2014).

22 Martino, D. J. *et al.* Longitudinal, genome-scale analysis of DNA methylation in twins from birth to 18 months of age reveals rapid epigenetic change in early life and pair-specific effects of discordance. *Genome Biology* **14**, R42 (2013).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

23 Martino, D. J. *et al.* Evidence for age-related and individual-specific changes in DNA methylation profile of mononuclear cells during early immune development in humans. *Epigenetics* **6**, 1085-1094 (2011).

24 Fitzmaurice, G. M., Laird, N. M. & Ware, J. H. *Applied longitudinal analysis*. Vol. 998 (John Wiley & Sons, 2012).

25 Ng, J. W. Y. *et al.* The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. *Genome Biology* **13**, 246 (2012).

26 Naumova, E. N., Must, A. & Laird, N. M. Tutorial in biostatistics: evaluating the impact of 'critical periods' in longitudinal studies of growth using piecewise mixed effects models. *International Journal of Epidemiology* **30**, 1332-1341 (2001).

27 Cudeck, R. & Harring, J. R. Analysis of nonlinear patterns of change with random coefficient models. *Annu. Rev. Psychol.* **58**, 615-637 (2007).

28 Xu, C.-J. *et al.* The emerging landscape of dynamic DNA methylation in early childhood. *Bmc Genomics* **18**, 25 (2017).

29 Richmond, R. C. *et al.* Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). *Human molecular genetics* **24**, 2201-2217 (2014).

30 Simpkin, A. J. *et al.* Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Human molecular genetics* **24**, 3752-3763 (2015).

31 Sharp, G. C. *et al.* Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* **44**, 1288-1304 (2015).

32 Ball, M. P. *et al.* Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. *Nature biotechnology* **27**, 361 (2009).

33 Wagner, J. R. *et al.* The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology* **15**, R37 (2014).

34 Illingworth, R. S. & Bird, A. P. CpG islands–'a rough guide'. *FEBS letters* **583**, 1713-1720 (2009).

35 Ong, C.-T. & Corces, V. G. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature Reviews Genetics* **12**, 283 (2011).

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

37 Bohlin, J. *et al.* Prediction of gestational age based on genome-wide differentially methylated regions. *Genome Biology* **17**, 207 (2016).

38 Hedman, Å. K. *et al.* Epigenetic patterns in blood associated with lipid traits predict incident coronary heart disease events and are enriched for results from genome-wide association studies. *Circulation: Cardiovascular Genetics* **10**, e001487 (2017).

39 Ligthart, S. *et al.* DNA methylation signatures of chronic low-grade inflammation are associated with complex diseases. *Genome Biology* **17**, 255 (2016).

40 Mendelson, M. M. *et al.* Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian randomization approach. *PLoS medicine* **14** (2017).

41 Demerath, E. W. *et al.* Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Human molecular genetics* **24**, 4464-4479 (2015).

42 Joehanes, R. *et al.* Epigenetic signatures of cigarette smoking. *Circulation: cardiovascular genetics* **9**, 436-447 (2016).

43 Sayols-Baixeras, S. *et al.* Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study. *Epigenetics* **10**, 1156-1165 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44 Sayols-Baixeras, S. *et al.* Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study. *Human molecular genetics* **25**, 4556-4565 (2016).

45 Singmann, P. *et al.* Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics & chromatin* **8**, 43 (2015).

46 Joubert, B. R. *et al.* DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *The American Journal of Human Genetics* **98**, 680-696 (2016).

47 Zeilinger, S. *et al.* Tobacco smoking leads to extensive genome-wide changes in DNA methylation. *PloS one* **8**, e63812 (2013).

48 Guida, F. *et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Human molecular genetics* **24**, 2349-2359 (2015).

49 Ambatipudi, S. *et al.* Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics* **8**, 599-618 (2016).

50 Chatziioannou, A. *et al.* Blood-based omic profiling supports female susceptibility to tobacco smoke-induced cardiovascular diseases. *Scientific reports* **7**, 42870 (2017).

51 Linnér, R. K. *et al.* An epigenome-wide association study meta-analysis of educational attainment. *Molecular Psychiatry* **22**, 1680-1690 (2017).

52 Zhu, X. *et al.* Genome-wide analysis of DNA methylation and cigarette smoking in a Chinese population. *Environmental health perspectives* **124**, 966-973 (2016).

53 Chen, W. *et al.* An epigenome-wide association study of total serum IgE in Hispanic children. *Journal of Allergy and Clinical Immunology* **140**, 571-577 (2017).

54 Yousefi, P. *et al.* Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *Bmc Genomics* **16**, 911 (2015).

55 Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human life course. *Genome research* **24**, 1725-1733 (2014).

56 Markunas, C. A. *et al.* Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. *Environmental health perspectives* **122**, 1147-1153 (2014).

57 Joubert, B. R. *et al.* Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nature Communications* **7**, 10577 (2016).

58 Marttila, S. *et al.* Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. *Bmc Genomics* **16**, 179 (2015).

59 Reynolds, L. M. *et al.* Age-related variations in the methylome associated with gene expression in human monocytes and T cells. *Nature Communications* **5**, 1-8 (2014).

60 Florath, I., Butterbach, K., Müller, H., Bewerunge-Hudler, M. & Brenner, H. Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Human molecular genetics* **23**, 1186-1201 (2014).

61 Li, S. *et al.* Identification, replication and characterization of epigenetic remodelling in the aging genome: a cross population analysis. *Scientific reports* **7**, 1-8 (2017).

62 Johansson, Å., Enroth, S. & Gyllensten, U. Continuous aging of the human DNA methylome throughout the human lifespan. *PloS one* **8**, e67378 (2013).

63 Felix, J. F. *et al.* Cohort profile: pregnancy and childhood epigenetics (PACE) Consortium. *International Journal of Epidemiology* **47**, 22-23u (2017).

64 Irizarry, R. A. *et al.* The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* **41**, 178 (2009).

65 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484 (2012).

66 Voigt, M., Hermanussen, M., Wittwer-Backofen, U., Fusch, C. & Hesse, V. Sex-specific differences in birth weight due to maternal smoking during pregnancy. *European Journal of Pediatrics* **165**, 757-761 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

67 Paus, T. *et al.* Corpus callosum in adolescent offspring exposed prenatally to maternal cigarette smoking. *Neuroimage* **40**, 435-441 (2008).

68 Toro, R. *et al.* Prenatal exposure to maternal cigarette smoking and the adolescent cerebral cortex. *Neuropsychopharmacology* **33**, 1019-1027 (2008).

69 Jacobsen, L. K., Slotkin, T. A., Mencl, W. E., Frost, S. J. & Pugh, K. R. Gender-specific effects of prenatal and adolescent exposure to tobacco smoke on auditory and visual attention. *Neuropsychopharmacology* **32**, 2453-2464 (2007).

70 Richmond, R. C., Suderman, M., Langdon, R., Relton, C. L. & Davey Smith, G. DNA methylation as a marker for prenatal smoke exposure in adults. *International Journal of Epidemiology* **47**, 1120-1130 (2018).

71 Kupers, L. K. *et al.* DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int J Epidemiol* (2015).

72 Nielsen, C. H., Larsen, A. & Nielsen, A. L. DNA methylation alterations in response to prenatal exposure of maternal cigarette smoking: A persistent epigenetic impact on health from maternal lifestyle? *Arch Toxicol*, doi:10.1007/s00204-014-1426-0 (2014).

73 Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics* **19**, 371 (2018).

74 Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology* **14**, 3156 (2013).

75 Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell* **49**, 359-367 (2013).

76 Heyn, H. *et al.* DNA methylation contributes to natural human variation. *Genome research* **23**, 1363-1372 (2013).

77 Barfield, R. T. *et al.* Accounting for population stratification in DNA methylation studies. *Genetic epidemiology* **38**, 231-241 (2014).

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

79 Gopalan, S. *et al.* Trends in DNA methylation with age replicate across diverse human populations. *Genetics* **206**, 1659-1674 (2017).

80 Kooijman, M. N. *et al.* The Generation R Study: design and cohort update 2017. *European journal of epidemiology* **31**, 1243-1264 (2016).

81 Boyd, A. *et al.* Cohort profile: the 'children of the 90s'—the index offspring of the Avon Longitudinal Study of Parents and Children. *International Journal of Epidemiology* **42**, 111-127 (2013).

82 Medina-Gomez, C. *et al.* Challenges in conducting genome-wide association studies in highly admixed multi-ethnic populations: the Generation R Study. *European journal of epidemiology* **30**, 317-330 (2015).

83 Relton, C. L. *et al.* Data resource profile: accessible resource for integrated epigenomic studies (ARIES). *International Journal of Epidemiology* **44**, 1181-1190 (2015).

84 Lehne, B. *et al.* A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biology* **16**, 37 (2015).

85 Min, J. L. *et al.* Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* (2018).

86 R Core Team. R: A language and environment for statistical computing. (2013).

87 Bakulski, K. M. *et al.* DNA methylation of cord blood cell types: applications for mixed cell birth studies. *Epigenetics* **11**, 354-362 (2016).

88 Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* **13**, 1 (2012).

89 Hothorn, T. *et al.* Package 'multcomp'. *Simultaneous inference in general parametric models. Project for Statistical Computing, Vienna, Austria* (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.09.142620; this version posted June 10, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

90 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823* (2014).

91 Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

92 Cecil, C. A. M. *et al.* Epigenetic signatures of childhood abuse and neglect: implications for psychiatric vulnerability. *Journal of Psychiatric Research* **83**, 184-194 (2016).