DNA co-methylation has a stable structure and is related to specific aspects of genome regulation

Sarah H Watkins*[1], Matthew Suderman[1], Gibran Hemani[1], Kimberly Burrows[1], Deborah A. Lawlor[1], Jane West[2], Kathryn Willan[2], Nicholas J Timpson[1], Josine L Min[1*], Tom R Gaunt[1*]

[1]MRC Integrative Epidemiology Unit, Department of Population Health Sciences, Bristol Medical School, University of Bristol, Bristol BS8 2BN, UK

[2] Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford Royal Infirmary, Duckworth Lane, Bradford, BD9 6RJ, UK

*Corresponding author

Email addresses: s.h.watkins@bristol.ac.uk; matthew.suderman@bristol.ac.uk; g.hemani@bristol.ac.uk; kimberly.burrows@bristol.ac.uk; d.a.lawlor@bristol.ac.uk; jane.west@bthft.nhs.uk; kathryn.willan@bthft.nhs.uk; n.j.timpson@bristol.ac.uk; josine.min@bristol.ac.uk; tom.gaunt@bristol.ac.uk

Abstract

Background

DNA methylation (DNAm) is an epigenetic mark that can be used to understand interindividual variability in genomic regulation, and it is influenced by many genetic and environmental factors. Co-methylation between DNAm sites is a known phenomenon but the full architecture of relationships between the approximately 450,000 (450k) sites most commonly measured in epidemiological studies has not been described. We investigate whether the *cis* and *trans* interindividual co-methylation structure amongst the 450k sites changes across the lifecourse, whether it differs between UK-born White and Pakistani individuals, and how it may be related to genome regulation.

Results

We find stability across the life course (birth to adolescence), across cohorts, and between two ethnic groups. Highly correlated DNAm sites in close proximity are highly heritable, influenced by nearby genetic variants (*cis* mQTLs), and are enriched for transcription factor (TF) binding sites related to regulation of short RNAs essential for cellular function transcribed by RNA polymerase III. Highly correlated sites that are either distant, or on different chromosomes, are driven by both common and unique environmental factors, and methylation at these sites is less likely to be driven by genotype. They are enriched for a multitude of TF binding sites and for inter-chromosomal chromatin contact sites, suggesting that DNA co-methylation of distant sites may be related to long-range cooperative TF interactions.

Conclusions

We conclude that DNA co-methylation, especially in *cis*, has a stable structure from birth to adolescence and between white British and Pakistani individuals. We hypothesise that co-methylation may have roles in genome regulation in humans, including 3D chromatin architecture.

The stable structure we have identified might have implications for the future design and interpretation of epigenetic studies.

Key words: DNA methylation, co-methylation, genome regulation, correlation, ALSPAC, Born in Bradford

Background

DNA methylation (DNAm) is an epigenetic modification to DNA that can be influenced by both environmental and genetic factors (1-3). It has roles in a variety of genomic functions, including a complex and context-dependent alteration of gene expression (4-7); both altering and being altered by transcription factor binding (8-13); and interaction with chromatin states (14-17). Epigenome wide association studies, where DNAm sites are tested in isolation, have identified DNAm sites associated with many exposures, diseases, and phenotypes (18-23). Networks of co-methylated DNAm sites have also been described - these utilise the relationships between DNAm sites to infer regulatory pathways in which DNAm might be involved (24-27). In contrast to the genetic architecture, our basic knowledge of relationships between DNAm sites is still limited, constraining our understanding of the biological mechanisms underlying co-methylation, and whether co-methylation has a role in genome regulation. Practically, co-methylation structure is an important consideration for the design of epigenome-wide association studies (EWAS) and differentially methylated region (DMR) studies; no equivalent of linkage disequilibrium (LD) matrices yet exists for DNAm sites. A deeper understanding of DNAm correlation structure, and its stability within and between individuals, might help identify whether this structure can be used for tagging, pruning and imputation in EWAS in a similar way to how LD enables these approaches to be used in genome-wide association studies (GWAS).

Both whole genome bisulfite sequencing (WGBS) and array-based studies highlighted that DNA methylation forms local correlation structures, with DNAm sites within 1-2kb often having correlated

62 methylation states (28-34); WGBS data shows that immediately adjacent sites almost always have
63 the same methylation state (28). Correlations between close DNAm sites (referred to here as *cis*
64 correlations) may be driven by genetic variants, as regions of highly correlating DNAm sites have
65 been associated with nearby SNPs (30, 33); although other studies have shown that correlations
66 between DNAm sites can be driven by environmental exposures (35). As the co-methylation
67 structure does not mirror the large blocks that linkage disequilibrium (LD) forms (30) and it appears
68 to be consistent across ethnic groups with different genetic architecture (32), it is unlikely to be
69 related to LD. There have been conflicting reports of whether DNAm correlation structure is related
70 to genomic annotations such as CpG islands (30, 34), but recent work has demonstrated that DNAm
71 sites within 2Mb at which methylation level co-varies due to the same causal genetic variants can be
72 used to predict contact sites for chromatin loops (36).

73 Co-methylation between distant DNAm sites on the same chromosome, and sites on different
74 chromosomes (referred to here as *trans* correlations), are less well described. Correlation of a
75 limited number of *trans*-chromosomal DNAm sites with highly variable DNAm levels showed that
76 DNAm sites related to *HOX* genes were highly correlated, suggesting *trans*-chromosomal correlations
77 might be related to biological pathways (35). Another recent paper using mouse tissue has shown
78 that DNAm sites around inter- and intra-chromosomal chromatin contact points have correlated
79 methylation states, with those within the same topologically associating domains and with the same
80 chromatin states having more correlated methylation states (16). This suggests correlations between
81 DNAm sites are likely to be relevant to genome regulation; however this has yet to be shown in
82 human studies for inter-chromosomal contacts.

83 DNAm is a dynamic epigenetic mark, so it is unsurprising that measurements at individual sites are
84 not always stable. Sites influenced by genotype tend to be more stable than those that are not; twin
85 models have found that the heritability of DNAm is on average 19% (1, 37, 38), with a subsequent
86 study showing the reliability of each of DNAm probe measurement is associated with the heritability

of the probe (39). DNAm changes with age (40-42) and in response to environmental exposures such as smoking (18, 43), but environmental and genetic constraints have been shown to contribute to the stability of DNAm over the lifecourse (37). As yet there is no indication of how stable relationships between DNAm sites might be; a lack of stability might indicate changing, non-essential functions, and would limit the utility of correlated methylation states in inferring genome regulation over different datasets. It is therefore important to assess correlation structure over time in the same individuals, and across datasets (to ensure replication). An important consideration when investigating stability is utilising datasets that include diverse social groups; DNA methylation is affected by adversity (44, 45), racial discrimination (46), social inequalities (19, 47), and environmental exposures such as air pollution (48, 49). Consequently, co-methylation structure is more likely to reflect stable biological processes if observations persist across diverse social groups (50).

In this paper, we outline DNA co-methylation structure in blood across the widely used Illumina 450k Beadchip which mainly covers promoters, TSS, and coding transcripts across 1.5% of the methylome (51, 52). We use data from two large UK birth cohorts – the Accessible Resource for Integrated Epigenomic studies (ARIES), which is a sub study of the Avon Longitudinal Study of Parents and Children (ALSPAC); and Born in Bradford (BiB). ARIES has longitudinal DNAm data from 849, 910, and 921 White British participants, at birth, 7 and 15-17 years, respectively; 788 participants have data at all three timepoints. BiB is a birth cohort recruited in a city with high levels of deprivation; the subcohort with DNAm data is approximately evenly split between two UK-born ethnic groups, with 424 white British and 439 Pakistani participants, measured at birth. In our analyses we split BiB by ethnic group (where membership of ethnic groups was defined by maternal self-report, and ethnicity is conceptualised as a social construct that is associated with differing social and environmental exposures, and therefore potentially differing effects on the methylome) to test whether co-methylation structure persists between two social groups, living within the same

113 geographic area and assayed on the same methylation array. We assess the stability of the co-
114 methylation structure between DNAm sites over time, across cohorts, and between two ethnic
115 groups (encompassing a wider range of social circumstances and enabling greater transportability of
116 our findings). We detail genetic and environmental influences on this correlation structure, and
117 provide the first comprehensive analysis of strong *cis* and *trans* co-methylation between DNAm sites
118 across the entire 450k array, outlining their distinct potential roles in genome regulation. We will
119 also provide a resource which can be used by the scientific community.

120 Results

121 Co-methylation structure and stability, across the 450k array

122 Overall co-methylation structure

123 To assess co-methylation structure, for each dataset we created a correlation matrix between all
124 possible pairs of sites on the Illumina HumanMethylation450 (450k) array in ARIES, and all sites on
125 the HumanMethylationEPIC (EPIC) array that also feature on the 450k array in BiB (see methods,
126 Figure 11). To assess features of correlations of different strengths, pairs were aggregated into
127 bands of correlation, ranging from -1 to 1, in increments of 0.1 (See Methods). For all 5 datasets, the
128 distribution of all possible pairwise correlations is positively skewed, and 83-87% of correlations are
129 between -0.2 and 0.2 (Figure 1; see supplementary figure 1 for the plot split by *cis* and *trans*
130 correlations). On average, just 12.6% of *cis* and 14.5% of *trans* pairs had a correlation above 0.2 or
131 below -0.2. To investigate whether physical proximity has an influence on the correlation, we
132 categorised the correlations in local and distant correlations defining *cis* as within 1Mb. The
133 proportions of all *cis* and *trans* correlations in each correlation band illustrate that physical proximity
134 between DNAm sites influences the likelihood of them having highly correlated methylation states
135 (Table 1). In all datasets, there are relatively equal percentages of *cis* and *trans* correlations for most
136 strengths of correlation. However, this differs substantially for the highest positive correlations (0.9-
137 1), where we see a higher proportion of *cis* than *trans* correlations (p=<2.2e-16 in all datasets using a

chi squared test). We also see a higher proportion of *trans* correlations in the BiB cohort than in ARIES (p=<2.2e-16 using a chi squared test at birth between ARIES and the BiB white British individuals), and we see a higher proportion of both *cis* and *trans* correlations in BiB for all correlation bands R>0.5 (p=<2.2e-16 using a chi squared test at birth between ARIES and the BiB white British individuals). As the higher correlations reflect less than 15% of all correlations, they can be viewed more clearly in Table 1 than Figure 1.

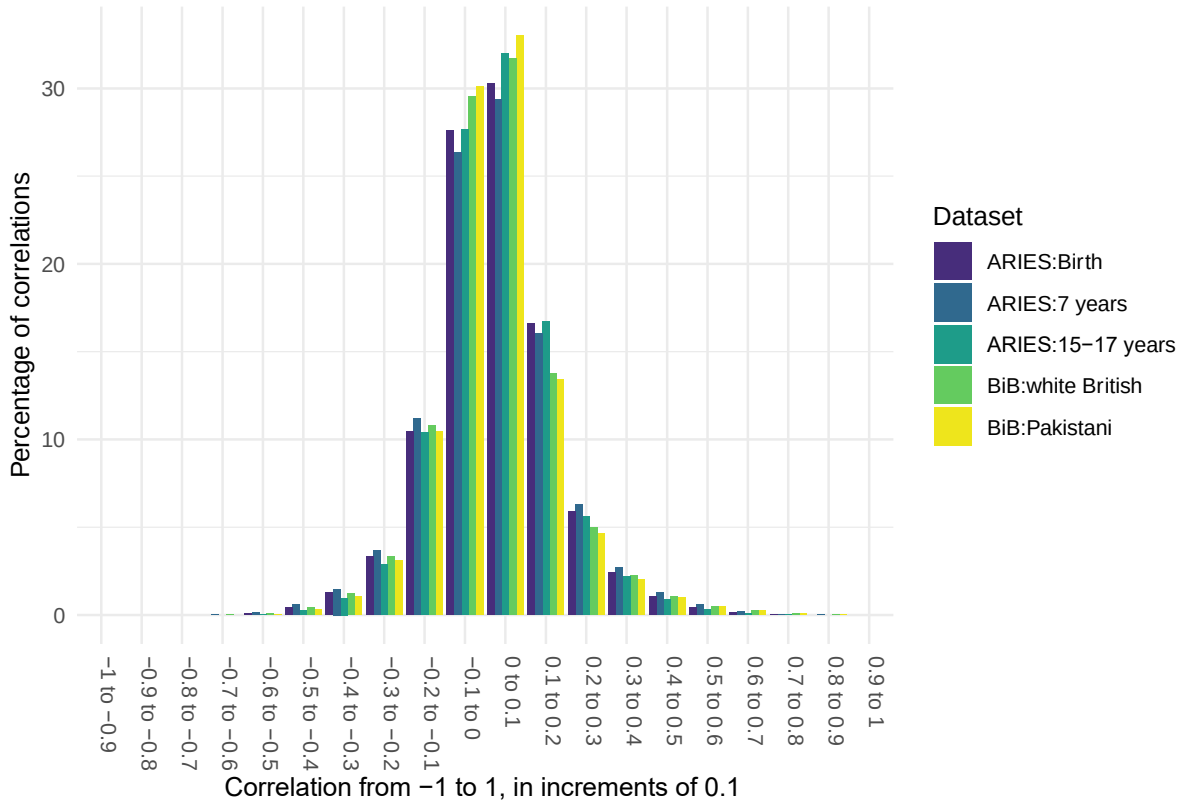

Figure 1: Bar plot showing the distribution of biweight mid-correlation values between all filtered sites on the 450k array

Table 1: Table of the numbers (percentages) of cis and trans biweight mid-correlations in each correlation band, comparing all ARIES and BiB datasets.

	ARIES Birth		**ARIES 7 years**		**ARIES 15-17 years**		**BiB Birth: White British**		**BiB Birth: Pakistani**	
Correlation band	Cis (total n= 119720281)	Trans (total n= 77830184780)	Cis (total n= 119720281)	Trans (total n= 77830184780)	Cis (total n= 119720281)	Trans (total n= 77830184780)	Cis (total n= 104564674)	Trans (total n= 68269791236)	Cis (total n= 104564674)	Trans (total n= 68269791236)
-1 to -0.9	0 (0)	3 (3.9e-09)	7 (5.8e-06)	0 (0)	7 (5.8e-06)	0 (0)	0 (0)	861 (1.3e-06)	0 (0)	298 (4.4e-07)
-0.9 to -0.8	3 (2.5e-06)	401 (5.2e-07)	35 (2.9e-05)	359 (4.6e-07)	35 (2.9e-05)	177 (2.3e-07)	45 (4.3e-05)	24088 (3.5e-05)	16 (1.5e-05)	9086 (1.3e-05)
-0.8 to -0.7	43 (3.6e-05)	4470 (5.7e-06)	116 (9.7e-05)	39097 (5.0e-05)	60 (5.0e-05)	1075 (1.4e-06)	403 (3.9e-04)	175848 (2.6e-04)	196 (1.9e-04)	80979 (1.2e-04)
-0.7 to -0.6	979 (8.2e-04)	922237 (0.001)	4742 (0.004)	4795746 (0.006)	322 (2.7e-04)	183597 (2.4e-04)	3261 (0.003)	2729487 (0.004)	1325 (0.001)	881459 (0.001)
-0.6 to -0.5	36468 (0.031)	42974963 (0.06)	81067 (0.07)	95777908 (0.1)	15467 (0.01)	19415165 (0.02)	48650 (0.05)	53148470 (0.08)	24738 (0.02)	26749688 (0.04)
-0.5 to -0.4	274527 (0.23)	317273367 (0.4)	404340 (0.3)	450807392 (0.6)	173905 (0.1)	206058410 (0.3)	267285 (0.3)	282045322 (0.4)	191497 (0.2)	208213229 (0.3)
-0.4 to -0.3	959877 (0.8)	1001068181 (1.3)	1171964 (1)	1150801019 (1.5)	748813 (0.6)	753206253 (1)	867102 (0.8)	828046988 (1.2)	742056 (0.7)	729079177 (1.1)
-0.3 to -0.2	2879403 (2.4)	2593874102 (3.3)	3345991 (2.8)	2880767119 (3.7)	2628050 (2.2)	2242288139 (2.9)	2673802 (2.6)	2270361950 (3.3)	2468830 (2.4)	2129015877 (3.1)
-0.2 to -0.1	10426258 (8.7)	8147044273 (10.5)	11178176 (9.3)	8693771577 (11.2)	10545834 (8.8)	8106743185 (10.4)	9759864 (9.3)	7355856656 (10.8)	9444571 (9)	7138202626 (10.5)
-0.1 to 0	30864827 (25.8)	21476134284 (27.6)	30021734 (25.1)	20523054551 (26.4)	30539452 (25.5)	21542953870 (27.7)	30546614 (29.2)	20177889631 (29.6)	31356436 (30)	20543387044 (30.1)
0 to 0.1	40717688 (34)	23557573193 (30.3)	39800719 (33.2)	22864354192 (29.4)	42748394 (35.7)	24869018033 (32)	36475305 (34.9)	21656131682 (31.7)	37728258 (36.1)	22534057813 (33)
0.1 to 0.2	22386616 (18.7)	12933163606 (16.6)	21062427 (17.6)	12473794995 (16)	21802894 (18.2)	12999434519 (16.7)	14832604 (14.2)	9391333794 (13.8)	14258103 (13.6)	9159570056 (13.4)
0.2 to 0.3	6863876 (5.7)	4596105049 (5.9)	7358153 (6.2)	4892148390 (6.3)	6555957 (5.5)	4359633908 (5.6)	4831645 (4.6)	3394420353 (5)	4447992 (4.3)	3163558256 (4.6)

0.3 to 0.4	2564609 (2.1)	1875571273 (2.4)	2997820 (2.5)	2120261423 (2.7)	2447776 (2)	1716228979 (2.2)	2188682 (2.1)	1521328435 (2.2)	1978938 (1.9)	1396655448 (2.1)
0.4 to 0.5	1070912 (0.9)	821068527 (1.1)	1349404 (1.1)	1011296118 (1.3)	969174 (0.8)	685508087 (0.9)	1074701 (1)	722291874 (1.1)	982620 (0.9)	665865317 (1)
0.5 to 0.6	460221 (0.38)	336608066 (0.4)	626390 (0.5)	467873540 (0.6)	371346 (0.3)	241537376 (0.3)	542477 (0.5)	349881373 (0.5)	506367 (0.5)	322746560 (0.5)
0.6 to 0.7	167909 (0.14)	106864972 (0.1)	247738 (0.2)	165454442 (0.2)	131295 (0.1)	70360685 (0.09)	296157 (0.3)	179320379 (0.3)	278815 (0.3)	167808304 (0.2)
0.7 to 0.8	41995 (0.04)	22537164 (0.03)	62118 (0.05)	32873193 (0.04)	37271 (0.03)	16648866 (0.02)	135469 (0.1)	75362451 (0.1)	131674 (0.1)	73361627 (0.1)
0.8 to 0.9	3832 (0.003)	1396572 (0.002)	6816 (0.006)	2312808 (0.003)	3814 (0.003)	964425 (0.001)	20326 (0.02)	9438017 (0.01)	21975 (0.02)	10543441 (0.02)
0.9 to 1	238 (2.0e-04)	77 (9.9e-08)	524 (4.4e-04)	911 (1.2e-06)	415 (3.5e-04)	31 (4.0e-08)	282 (2.7e-04)	3577 (5.2e-06)	267 (2.6e-04)	4951 (7.3e-06)

Cis co-methylation structure across the genome (as measured by the 450k)

To investigate the influence of physical distance on the *cis* co-methylation structure in more detail, decay plots were created across all autosomes (as measured by the 450k array) separating positive and negative correlations. In line with previous work, *cis* correlations within 10kb across the whole genome reveals a smooth decay that reduces to background correlation (about 0.125) at approximately 3kb (28, 30, 34) and the decay is identical across population cohorts and across ethnic groups (32). This suggests that it is unlikely to be driven by LD, partly because the decay is over a vastly smaller distance (see LD decay plot in (53) for comparison), and partly because the decay is identical for two different ethnic groups. Furthermore, the decay shows a high degree of heterogeneity and is identical between birth and adolescence, meaning *cis* correlation does not solely depend on physical distance (Figure 2, supplementary figure 2). In contrast to positive correlations, we also found for the first time that negative correlations are not distance dependent as immediately adjacent sites show the lowest correlations and have a slight peak between 2.5 and 4kb.

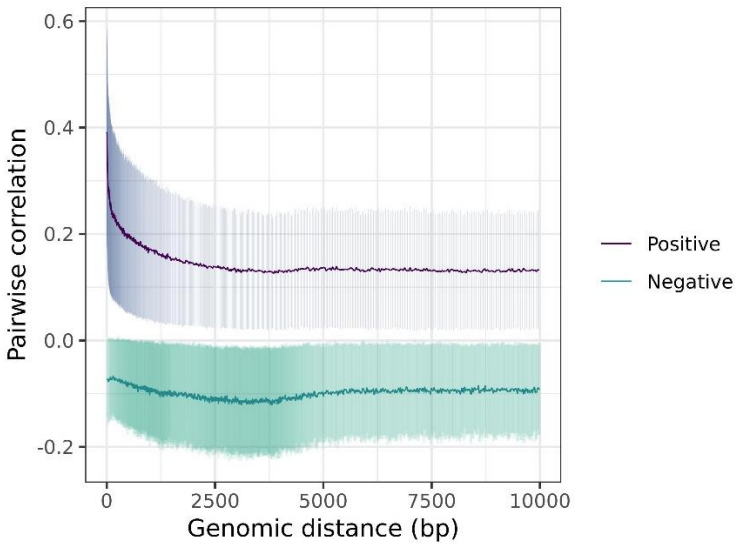

Figure 2: Decay plots of pairwise cis correlations (n=3,114,257, calculated using the biweight mid-correlation) from all filtered sites within 10kb of each other on the 450k array, across all autosomes (in ARIES 7 year olds;

plots for all datasets can be found in supplementary figure 2). The variance in each bin (represented here by the bin standard deviation) was added to the plot to demonstrate the heterogeneity around the binned estimates.

Stability of correlations across time, datasets, and ethnic groups

To assess whether strong correlations between DNAm sites (R>0.8, n=2,483,055 to 13,145,092) differ across time, cohorts, and ethnic groups more than expected by chance, we plotted mean difference (i.e. for each CpG plotting mean correlation vs difference in correlation between two groups). Based on 95% confidence intervals, we do not find strong evidence of a difference in the number of either *cis* or *trans* correlations between any of our datasets, suggesting that correlations R>0.8 between DNAm sites are relatively stable (see **Error! Reference source not found.**). There are smaller mean changes in correlation between birth and 7 years (-0.015 for *cis*; 0.006 for *trans*) than between 7 years and adolescence (0.023 for *cis*; 0.028 for *trans*), indicating that changes may increase with age; specifically that correlations reduce with age (with the exception of *cis* correlations between birth and 7 years). Although the 95% confidence intervals include 0, correlations are consistently higher in the BiB white British individuals compared to ARIES at birth (0.085 for *cis*; 0.088 for *trans*), which is reflected in the higher proportion of strong correlations in BiB. In contrast, the mean difference between BiB white British and Pakistani groups is very small (0.003 for *cis*; 0.005 for *trans*), suggesting that between-cohort differences are stronger than between ethnic groups. Supplementary figure 3 shows that there are a small number of DNAm sites at which there are large changes in *trans* correlation between birth and 7 years in ARIES, between ARIES and BiB white British individuals, and between the two ethnic groups in BiB.

Table 2: Mean differences in correlations (R>0.8, n=2,483,055 to 13,145,092) plus SD and 95% confidence intervals (CI) between timepoints and datasets, for cis and trans correlations

	Mean (cis)	**SD (cis)**	**95% CI**	**Mean (trans)**	**SD (trans)**	**95% CI**
ARIES vs BiB (white British)	0.085	0.07	-0.05 to 0.22	0.088	0.06	-0.03 to 0.2
BiB (white British vs Pakistani)	0.003	0.03	-0.06 to 0.07	0.005	0.03	-0.05 to 0.06
ARIES birth vs 7	-0.015	0.05	-0.11 to 0.08	0.006	0.03	-0.04 to 0.06
ARIES 7 vs 15-17	0.02	0.03	-0.03 to 0.07	0.03	0.03	-0.02 to 0.08

Genetic and environmental influences on co-methylation

Influence of genetic factors on co-methylation

To assess whether the correlation strengths are influenced by genetic factors, we used twin heritabilities of DNAm sites (1) and assessed the proportion of correlations which had zero, one, or two of the DNAm sites in each correlating pair associated with an mQTL (see Methods). The heritability of both sites in highly correlating pairs (r>0.9) tend to be matched, suggesting that co-methylation is related to heritability in *cis* but not in *trans* (see Figure 3B). In contrast to sites with lower correlations, *cis* sites correlating >0.9 are highly heritable (with a mean contribution of heritability to variance of 0.77 across all datasets), and at least 84% of *cis* correlations >0.9 have both DNAm sites associated with a *cis* mQTL across the 5 datasets (see Figure 3 and supplementary figure 4).

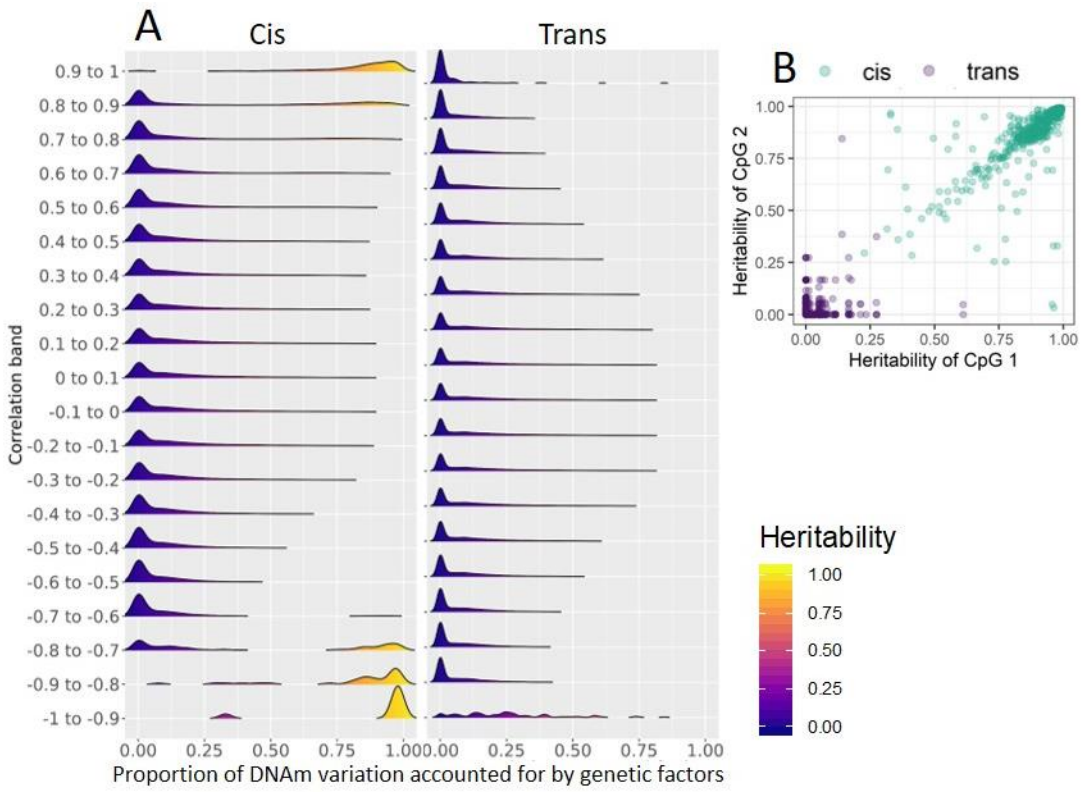

Figure 3: ***A****: Density plots illustrating the proportion of DNAm variation due to heritability for DNAm sites with correlations of differing strengths, in ARIES 7 year olds (please see supplementary figure 4 for plots of each dataset).* ***B****: Scatter plot of heritability of probe pairs correlating >0.9 in ARIES 7 year olds.*

To identify whether co-methylation is influenced by mQTL, we firstly assessed the proportion of correlations which had zero, one, or two of the DNAm sites in each correlating pair associated with an mQTL. At least 84% of *cis* correlations >0.9 have both DNAm sites associated with a *cis* mQTL across the 5 datasets, in line with the heritability findings above. *Trans* correlations >0.9 are most likely to have neither DNAm site associated with a *cis* mQTL in 4 of the datasets (48-55%; in the BiB white British group only 31% are associated with 0 mQTLs). This suggests that strong *trans* correlations are between sites that are less likely to have been associated with an mQTL (although it is possible they have shared genetic aetiology we have not detected - *trans* sites are more polygenic and there are fewer detected *trans* mQTLs at present (54, 55)). There are slight increases in the proportion of *cis* correlations that have both DNAm sites associated with an mQTL in the ARIES

adolescents. This is noticeable for *cis* correlations between -0.8 and -0.6, and between 0.8 and 0.9 (see Figure 4 and supplementary figure 5).

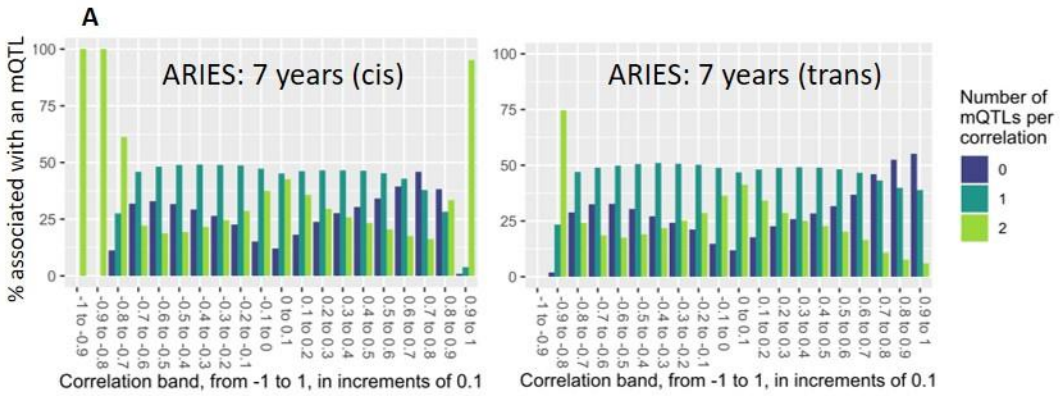

*Figure 4: (**A**): Bar plots of the percentage of pairwise correlations in each correlation range that have 0, 1 or 2 DNAm sites associated with an mQTL identified by the GoDMC consortium. Split by cis (left) and trans (right) correlating pairs, In ARIES at 7 years.*

Influence of environment on co-methylation

Environmental influences on DNAm have been separated as common (or shared) environment and unique environment which includes measurement error (1). Consistent with the high heritability estimates for sites that correlate r>0.9 in *cis*, the common and unique environment estimates are low. In *trans,* almost all sites correlating r>0.9 are influenced on average 41% by common environment and 49% by unique environment (measurement error and interindividual stochastic variation).

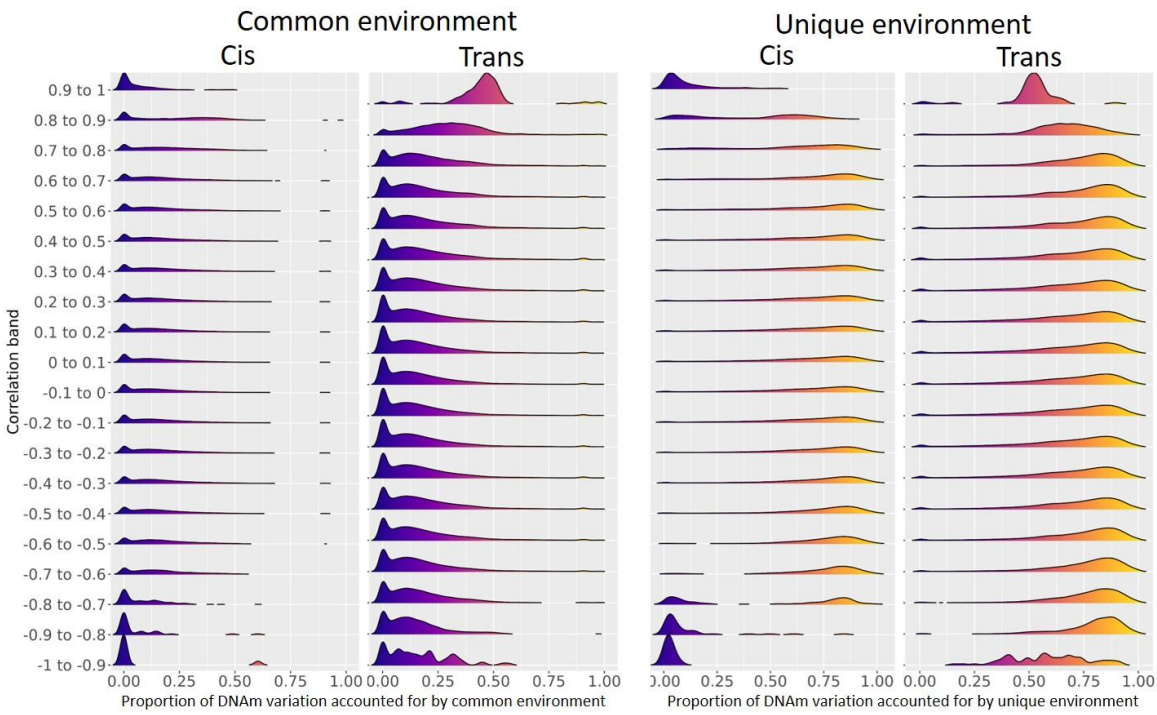

Figure 5: Density plots illustrating the proportion of DNAm variation due to common environment and unique environment for DNAm sites with correlations of differing strengths, in ARIES 7 year olds (please see supplementary figure 6 for plots of each dataset).

Features of highly co-methylated sites

Overview

Because *cis* and *trans* sites correlated >0.9 have distinct features in terms of genetic and environmental influences, we conducted an in-depth analysis to identify the biological mechanism of this co-methylation. The number of correlations R>0.9 in each dataset are detailed in Table 3). Throughout the following section for simplicity we show results for the ARIES 7 year olds, as results were very similar across all five datasets (See Supplement for other datasets).

Table 3: Numbers of cis and trans correlations r>0.9, and the number of unique probes these correlations are between.

Cohort	Timepoint	*Cis* correlations	*Trans* correlations
ARIES	Birth (n=849)	238 (305 probes)	77 (46 probes)
	7 years (n=910)	524 (589 probes)	911 (162 probes)
	15-17 years (n=921)	415 (486 probes)	31 (24 probes)
BiB	Birth – white British (n=424)	282 (413 probes)	3577 (683 probes)
	Birth - Pakistani (n=439)	267 (387 probes)	4951 (669 probes)

Biological mechanisms of highly co-methylated sites

Cis and *trans* sites that have correlations >0.9 display very different patterns of variance. *Cis* sites are more variable in both mean methylation level and variance, although there is a tendency for hypo- and hyper- methylated sites to have smaller variance. *Trans* correlated sites are typically hypomethylated, and have low variance. As one might expect DNAm sites under *trans*-acting influences such as transcription factors to be hypomethylated (11, 56), this would fit the behaviour of the highly trans-correlated sites seen here (see Figure 6 for illustration).

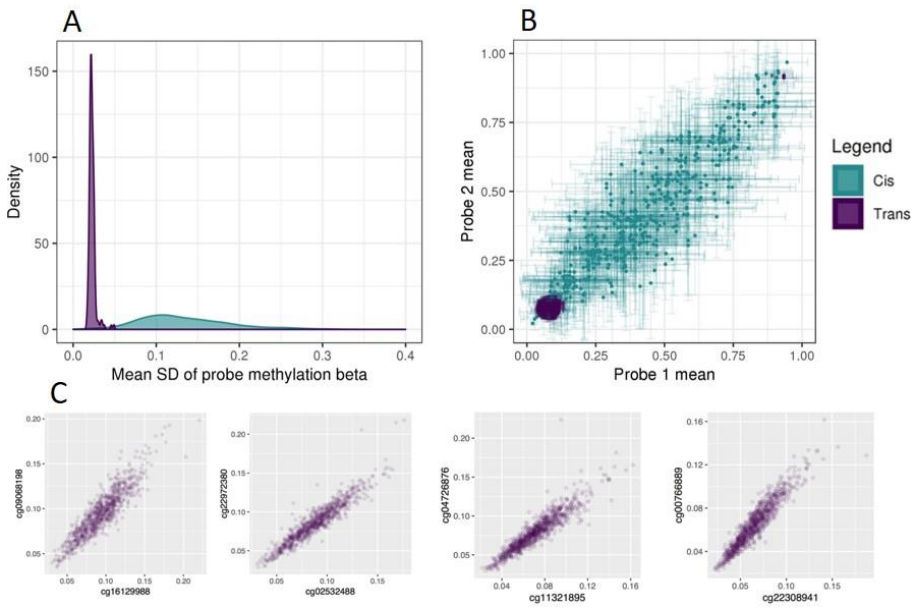

Figure 6: ***A****: Density plot showing the mean standard deviation of cis and trans DNAm sites with correlations R>0.9.* ***B****: Scatter plot with error bars (standard deviation) illustrating mean methylation level of cis and trans probe pairs that correlate R>0.9.* ***C****: Illustration of raw scatter plots of trans-correlating probe pairs R>0.9.*

Trans co-methylation structure

The architecture of co-methylation between sites that are either distant or on different chromosomes has not previously been well characterised. We find that in all datasets these sites are distributed across the genome and interconnected (illustrated by the circos and network plots in Figure 7 and supplementary figures 8 and 9). There is a much higher number of correlations r>0.9 in BiB than in ARIES (as shown in Table 1 and the section on overall co-methylation structure). This can be explained by correlations in BiB being on average 0.09 higher than in ARIES (Table 2); something that may be due to array or sample type effects. The connections between the sites correlating r>0.9 resemble scale-free networks (57, 58) with a small number of 'hub' nodes having large numbers of connections. All datasets except 15-17 years have a power-law alpha between 2 and 3, which is indicative of a scale-free network (57, 59) (however at 15-17 years the number of nodes is too small to assess scale-free properties appropriately). Degree distribution plots are found in supplementary figure 7.

To test the preservation of network architecture for DNAm sites correlating in *trans* >0.9, we tested preservation of both the DNAm sites (nodes) and the connections between them (edges) between the five datasets. To do this we utilised the Cytoscape Network Analyzer to test the intersection between pairs of networks. Using the Binomial test we find strong evidence of preservation of both the nodes and the edges of the networks for all four comparisons; between ARIES and BiB white British cord blood, between BiB white British and Pakistani cord blood, between birth and 7 years in ARIES, and between 7 years and 15-17 years in ARIES. Results are detailed in Table 4.

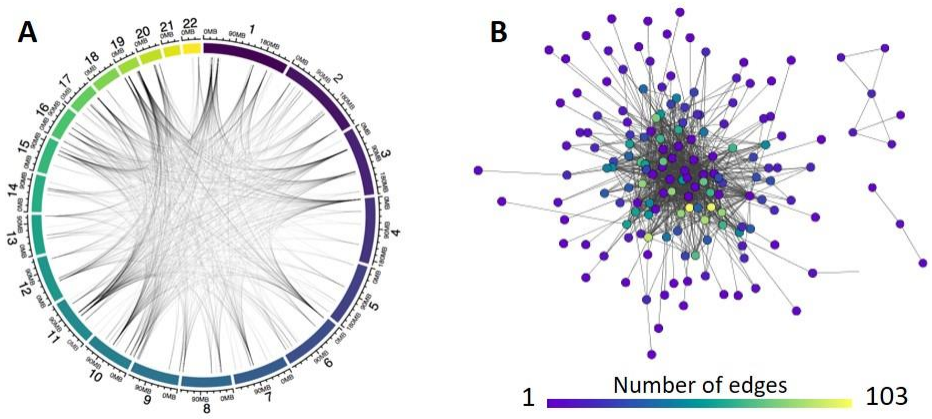

Figure 7: ***A****: circos plot illustrating genomic distribution of trans correlations r>0.9 in ARIES 7 year olds (circos plots for all datasets can be found in supplementary figure 8).* ***B****: cytoscape network plot illustrating network connectivity of trans correlations r>0.9 in ARIES 7 year olds (cytoscape network plots for all datasets can be found in supplementary figure 9).*

Table 4: Node and edge intersections for trans-correlating DNAm sites r>0.9. Nodes are DNAm sites and edges are the specific pairwise correlations between two nodes. Intersection means the node, or the specific correlation between two nodes, is present in both datasets A and B. N = number of intersections; p-values assess the likelihood of the overlap using the Binomial exact test.

Dataset A	**Dataset B**	**Dataset A nodes (edges) R>0.9**	**Dataset B nodes (edges) R>0.9**	**Intersecting nodes N (p-value)**	**Intersecting edges N (p-value)**
ARIES at birth	BiB white British at birth	46 (77)	683 (3577)	31 (7.7e-39)	9 (1.3e-04)
BiB white British at birth	BiB Pakistani at birth	683 (3577)	669 (4951)	401 (0)	1056 (0)
ARIES at birth	ARIES 7 year olds	46 (77)	162 (911)	35 (8.6e-47)	23 (9.3e-19)
ARIES 7 year olds	ARIES 15-17 year olds	162 (911)	24 (31)	21 (4.3e-31)	15 (1.9e-16)

Enrichments of *cis* co-methylated sites

To illustrate the potential utility of strong *cis* co-methylation, enrichment analyses were conducted to assess the enrichment for DNAm sites being located at specific transcription factor binding sites (TFBS), chromatin states, and genomic regions. All pairs of DNAm sites correlating R>0.9 and within 1Mb of each other were included in this analysis. Enrichments were virtually identical for all 5 datasets. *Cis* correlating sites are strongly enriched for chromatin states associated with poised promoters (PromP; OR=3.9 to 4.8, p=1.1e-09 to 2.4e-17), and less strongly enriched for weak enhancers (EnhW1 (OR=2.1 to 2.7, p=1.4e-03 to 8.4e-05), EnhW2 (OR=1.9 to 2.2, p=0.03 to 0.02). *Cis* correlating DNAm sites are strongly enriched for a select few TFBS in blood: RNA Polymerase III (Pol3) (OR=10.3 to 18.7, p=2.4e-05 to 4.6e-09), BRF1 (OR=10.5 to 18.5, p=2.9e-05 to 5e-09), and BDP1 (OR=6.8 to 11.4, p=2e-04 to 1.5e-06), with weaker enrichment of TFIIIC-110 and RPC155. Heatmaps of TFBS and chromatin enrichments are shown in Figure 8. *Cis* sites showed strongest enrichment for location in promoters (OR=1.62 to 2.5, p=6.2e-06 to 1.1e-17) and 5'UTR (OR=1.6 to 2, p=0.02 to 9.8e-09), with weaker enrichment in exons (OR=1.4 to 1.6, p=0.02 to 9.7e-06) (shown in supplementary figure 10). As BRF1 and BDP1 are essential for RNA polymerase III transcription, this

suggests coordination of methylation state of sites in close proximity is primarily a feature of active promoter regions involved in regulation of short RNAs essential for cellular function transcribed by RNA polymerase III.

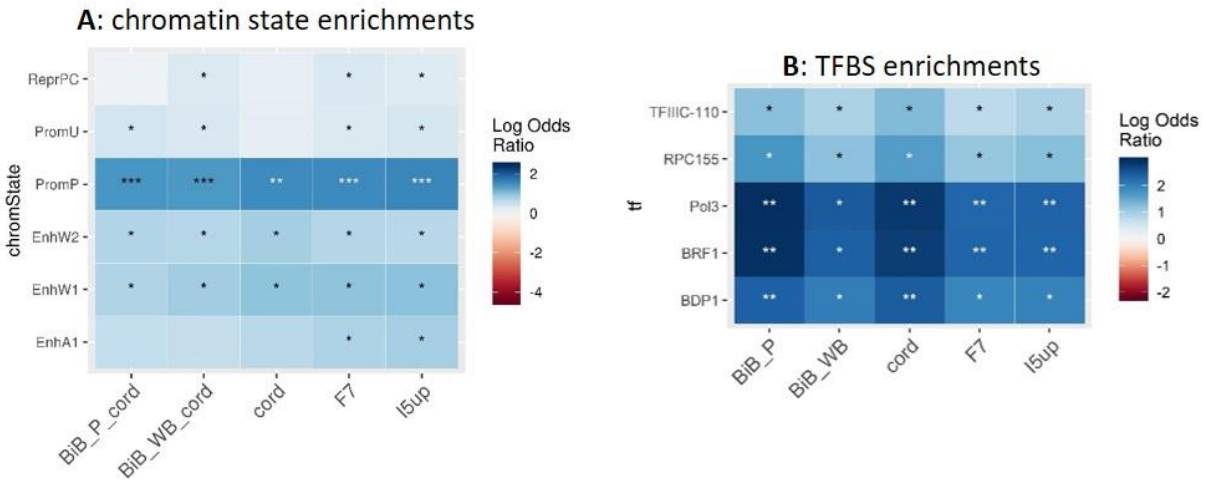

*Figure 8: Enrichments of cis-correlating sites r>0.9 for **A** chromatin states and **B** transcription factor binding sites (TFBS). Enrichment analyses were conducted using LOLA, with enrichments in blood only.; *p<0.05 **p<1e-05 ***p<1e-10*

Enrichments of *trans* co-methylated sites

Next, we tested whether sites involved in strong (R>0.9) *trans* correlations were enriched for locations in TFBS, chromatin states and genomic regions. *Trans* correlating sites are enriched for 29 of the tested TFBS: including Pol2 (OR=3.1 to 6.5, p=2.9e-03 to 5.2e-12), PHF8 (OR=3.6 to 6.4,p=3.2e-05 to 9.9e-81), MAZ (OR=3.3 to 5.7,p=9.3e-05 to 8.9e-63), ELF1 (OR=3 to 5,p=5.1e-04 to 3.6e-63), Egr-1 (OR=3.1 to 5.6, p=3e-04 to 2.5e-49), and TAF1 (OR=2.7 to 5.6, p=1.8e-03 to 9.5e-47). In contrast to *trans*-mQTL associated probes (54), we do not see enrichment for cohesin related TFs, which may reflect the small proportion of *trans* correlations found on the same chromosome. *Trans*-correlating sites are strongly enriched for chromatin states associated with promotors downstream of transcription start sites 1 (PromD1) (OR=4.1 to 11, p=5.1e-07 to 1.9e-78), and active transcription start sites (TssA) (OR=2.6 to 3.4, p=4.9e-02 to 1e-30), and weakly enriched for locations at promotors upstream of transcription start sites (PromU) (OR=1.6 to 2, p=0.26 to 1.8e-09). *Trans* sites show no

evidence of enrichments in the adolescents in ARIES; for all other datasets we see the strongest enrichment for CpG islands (OR=2.2 to 6.1, p=3.8e-08 to 5.7e-28), with moderate enrichment in promoters (OR=1.9 to 2.6, p=4e-04 to 1.3e-32) and 5'UTR (OR=2.1 to 2.2, p=3.3e-05 to 4.7e-21) (shown in supplementary figure 11). As such, co-methylation of *trans* sites may relate to active transcription.

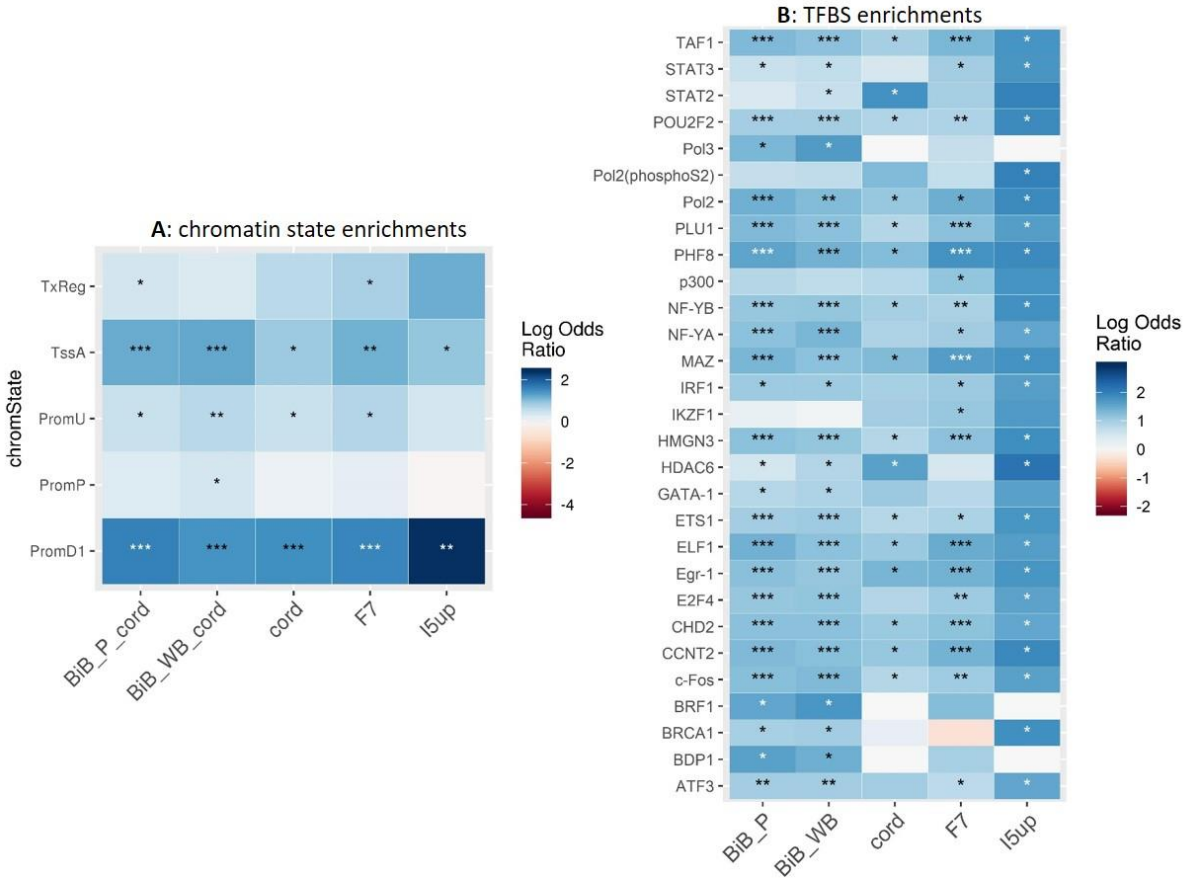

*Figure 9: Enrichments of trans-correlating sites r>0.9 for **A** chromatin states and **B** transcription factor binding sites (TFBS). Enrichment analyses were conducted using LOLA, with enrichments in blood only.; *p<0.05 **p<1e-05 ***p<1e-10*

Enrichment of inter-chromosomal chromatin contacts in *trans* co-methylated sites

To identify whether strong *trans* correlations are located in sites where chromatin contacts are formed, we assessed the overlap of the *trans* correlations r>0.9 with the (Rao et al., 2014) Hi-C data. This analysis was performed with the 860 inter-chromosomal *trans* correlations r>0.9 in the ARIES 7

year olds, and the 3577 in the BiB white British dataset. We find a strong enrichment of inter-chromosomal Hi-C contacts in the real correlation data as compared to 1000 permutations of the data, with no permutation set having a higher count of overlaps than the real data in either ARIES or BiB (n=46 in ARIES and n=654 in BiB; p=0 for both datasets) (Figure 10). This suggests that correlation of DNAm sites across chromosomes is at least in part likely to be related to inter-chromosomal contacts, which would mean that coordinated methylation states between DNAm sites on different chromosomes could be functionally relevant to inter-chromosomal chromatin contacts.

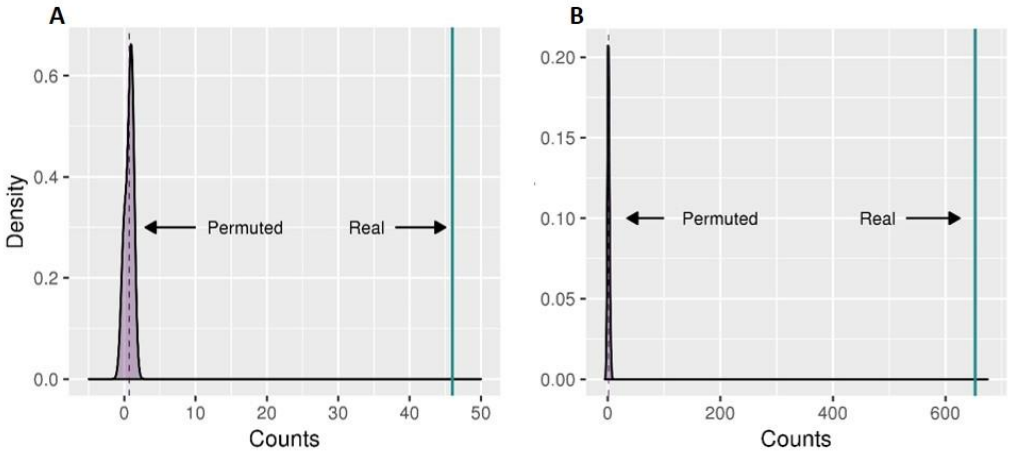

Figure 10: ***A****: In ARIES the overlap in the permuted datasets is never more than 1, whereas there are 46 overlaps in the real data.* ***B****: In BiB the overlap in the permuted datasets is never greater than 5, whereas there are 654 overlaps in the real data.*

Discussion

This work has demonstrated that DNAm data has a stable co-methylation structure, both in *cis* and in *trans*, that persists over time and is similar across datasets and two ethnic groups. This work suggests that co-methylation of DNAm sites may be related to specific aspects of genome regulation, where sites in *cis* are fundamentally different to co-methylation of sites which are distant or on different chromosomes. Through enrichment analyses we find that co-methylation is likely to be related to short RNA transcription associated with RNA polymerase III in *cis*, and multiple

transcription factors in *trans*. We have demonstrated in humans that *trans*-correlating DNAm sites are likely to represent inter-chromosomal contacts, and thus they are likely to represent shared regulation. This is consistent with recent published evidence of correlation between DNAm sites in inter- and intra- chromosomal chromatin contact regions in mice (16).

It is striking that all the birth datasets (one in ARIES and two in BiB) have very similar numbers of *cis* correlations >0.9. Numbers of *trans* correlations >0.9 are vastly different, with the BiB participants having substantially more. This is expected from our analysis of overall correlation structure, and our test of stability between cohorts (where correlations in BiB are higher than in ARIES by about 0.09, although this difference does not pass the 95% confidence intervals). This could be due to a number of factors; the differing array platforms, differing sample types (blood spots, white cells or whole blood in ARIES vs whole blood only in BiB), unaccounted for batch effects, or power (with ARIES timepoints being around twice the size of the two ethnic groups in BiB).

We find that co-methylation of sites R>0.8 does not differ between our datasets more than expected by chance, although the difference does increase with age in ARIES, suggesting that correlations may reduce over time (which one might expect from increasing stochastic variation). There is greater similarity between ethnic groups than between cohorts, which illustrates the importance of multi-ethnic studies (i.e., so we do not erroneously attribute differences in DNAm induced by cohort or batch effects to differences between ethnic groups). Stability of co-methylation over time, cohorts, and ethnic groups is important for two reasons. Firstly, the stability means that the sites identified in this study are likely to be highly correlated in other datasets – this gives our findings potential for broader applicability of co-methylation structure to EWAS and DMR analyses. Secondly, it suggests that strong co-methylation is likely to be related to basic biological functions that are not cohort or age dependent, and the consistency between ethnic groups means it is stable in the presence of social and genetic diversity (50) – this is important when considering the generalisability of our results.

We demonstrate that although *cis* co-methylation is distance-based, there is a large degree of variation. This is likely to reflect the fact that we find strong *cis* correlations are related to genomic regions (most strongly promotors and 5'UTRs). Strong *cis* correlations appear to be largely driven by genotype, as they are highly heritable and associated with *cis* mQTLs. From our enrichment analysis it appears that *cis* co-methylation, driven by genotype, is enriched for being located only at binding sites of transcription factors essential for RNA polymerase III transcription; this is supported by the recent demonstration that *cis* mQTL SNPs overlap TFBS (7, 54). Their coordinated DNAm states are therefore likely to be involved in regulation of the transcription of short RNAs involved in essential cellular functions (such as protein synthesis and transport) (60, 61), and, most relevant to the datasets we use in this study, RNA polymerase III transcription has been shown to be a determinant of growth of both the cell and the organism (62).

Strong correlations between DNAm sites on different chromosomes are shown here to be strongly enriched for inter-chromosomal contacts published in two independent datasets. Coordinated methylation states have been shown previously between contacting inter-chromosomal regions of the genome in mice (16). We demonstrate that this phenomenon can be identified using existing microarray and Hi-C data in humans. This is very much in line with recent work which shows co-methylated DNAm sites on the same chromosome are enriched for chromosomal loop contact sites (36). One might in fact expect the true number of inter-chromosomal contacts between highly co-methylated inter-chromosomal DNAm sites to be higher than reported here, as methods such as Hi-C do not pick up many inter-chromosomal contacts due to the greater distance between them than between *cis* contact sites (63). The enrichment for a multitude of transcription factor binding sites suggests inter-chromosomal co-methylation may be related to transcription factor networks, which have key roles in genome regulation (64) and have been shown to mediate inter-chromosomal chromatin contacts (65). Finally, highly co-methylated *trans* sites are influenced almost entirely by non-genetic factors – that makes co-methylation an interesting factor to study in relation to

environmental influences on genome regulation, given the chromatin and transcription factor enrichments for these sites.

There are of course some limitations to this work. The first is that our datasets only span birth to adolescence; we do not know how applicable our results will be to adults, although it might be reasonable to assume that correlations will decrease further. Differences between cohorts are likely to have been increased by the use of different array platforms (450k in ARIES vs EPIC in BiB), smaller numbers of BiB participants, and diversity of sample types in ARIES. This study only considered DNA methylation in blood - DNAm is cell-type- and tissue- specific, and as such the applicability of these results to DNAm in other tissues may be fairly limited (Hannon et al., 2015); although it is likely to be better for the highly heritable strong *cis* correlations (Hannon 2018). Of particular importance is a demonstration that *trans*-chromosomal promoter contacts have been shown to be cell-type specific (66), and so future work may benefit from utilisation of less heterogeneous cell type populations than blood or advanced deconvolution methods to investigate the functions of trans-chromosomal co-methylation. We cannot discount that the enrichments we see could be down to technical effects with probes in specific locations, or that the enrichments are due to as yet unknown genomic factors; we controlled for this as far as we could, using CG content matched probes from the 450k as background and removed probes that have been flagged as unreliable (67). Finally, this work cannot illustrate the exact role that co-methylation plays in these aspects of genome regulation – is co-methylation actively involved in regulation of transcription factor binding, for example, or is it a passive consequence? This is particularly interesting in the context of inter-chromosomal contacts, because we have identified that these DNAm sites are influenced by the environment, and so could provide a mechanism by which the environment affects genome regulation.

Conclusions

DNAm has a stable co-methylation structure in humans that persists to at least adolescence and across social groups (here represented by both cohort and ethnicity). *Cis* co-methylation is likely to

be related to short RNA transcription that is associated with RNA polymerase III. *Trans* co-methylation is highly enriched in regions of inter-chromosomal contacts, and for the binding sites of multiple transcription factors, suggesting that co-methylation may have a role in 3D genome regulation.

Materials and methods

Code availability

Scripts to replicate these analyses can be found on GitHub:

https://github.com/shwatkins/PhD/tree/master/450k_correlation_analysis

Data availability

ALSPAC: Data are available to researchers by request from the Avon Longitudinal Study of Parents and Children Executive Committee (http://www.bristol.ac.uk/alspac/researchers/access/) as outlined in the study's access policy http://www.bristol.ac.uk/media-library/sites/alspac/documents/researchers/data-access/ALSPAC_Access_Policy.pdf

BiB: Data are available to researchers who submit an expression of interest to the Born in Bradford Executive Group who review applications monthly and aim to respond within eight weeks. More details of data available and how to apply for access are on the BiB website: https://borninbradford.nhs.uk/research/.

Participants

Data were taken from participants of two birth cohorts: the Avon Longitudinal Study of Parents and Children (ALSPAC), and the Born in Bradford study (BiB). Detailed descriptions of the cohorts can be found in the supplementary methods; participant numbers are below in Table 5. In brief, ALSPAC is a multi-generational cohort study based in Bristol, comprising mostly white British participants. A subsample of participants (known as ARIES) of 1022 mother-child pairs had DNAm data generated at five timepoints: birth, 7 years and 15-17 years in the children, and during pregnancy and 12-18 years

later in the mothers. Our study utilises the three child timepoints only. BiB is a longitudinal, multi-ethnic cohort study based in Bradford, UK. Like ALSPAC it was set up to investigate factors which influence child health and development, but with a particular focus on child morbidity and mortality, as rates of these have been higher in Bradford than the rest of the UK (68). Bradford has a high rate of economic deprivation – one third of the neighbourhoods in Bradford are in the most deprived 10% of neighbourhoods in England (69, 70). Around 20% of the population are of South Asian descent, and 90% of these individuals are of Pakistani origin (68); the BiB DNAm subsample was specifically designed to be multi-ethnic, and so of those eligible, 500 White British and 500 Pakistani mothers were selected to have DNAm generated for themselves and their children.

Participant numbers

Table 5: Overview of study participants and sociodemographic characteristics

	ARIES			**BiB**	
Age	Birth	7 years	15-17 years	Birth	Birth
Self-reported race/ethnicity	White British	White British	White British	White British	Pakistani
Number of participants	849	910	921	424	439
% female	52%	51%	51%	48%	48%
Blood sample type	Blood spots, white cells	White cells, whole blood	White cells	Whole blood	Whole blood
Household class % high	61	60	61	NA	NA
Household class % medium	31	31	30	NA	NA
Household class % low	4	4	4	NA	NA
Index of Multiple Deprivation mean(SD)	NA	NA	NA	36.9 (19.7)	46.9 (14.8)

DNA methylation data

DNA methylation data generation is described in detail for each cohort in the supplementary methods.

ARIES

Briefly, in ARIES consent for biological samples was collected in accordance with the Human Tissue Act (2004). Processing, extraction, and quality control of DNAm data has been described in detail for these samples (71), as have normalisation and outlier removal procedures (72). We removed 21 further individuals as they were the only sample on a slide, preventing their adjustment for slide effects. This left us with 849 DNAm profiles at birth, 910 at 7 years, and 921 at 15-17 years.

BiB

Consent for biological samples was collected in accordance with the Human Tissue Act (2004). To generate DNAm data, the EZ-96 DNA methylation kit (Zymo Research, Orange, CA, USA) was used to bisulfite-convert 500ng of high molecular weight DNA. DNAm was assessed using the Illumina Infinium MethylationEPIC beadchip arrays (Illumina, San Diego, CA, USA). Batch variables were recorded using a laboratory information management system (LIMS). The Beta-value statistic was used to represent methylation levels. 2010 samples (mothers, children and control samples) were generated together. 98 failed genotype concordance checks, and 13 failed other QC checks, and as some of these issues overlapped in samples from the same individuals, a total of 100 individuals were removed during QC. This left 1910 samples for normalisation, comprising 25 controls, 934 mother and 951 child participants. For this study we used DNAm data only for the children; 88 further participants were removed as they were related >12.5%.

DNA methylation data normalisation

A detailed description of normalisation methods and site filters are in the supplementary methods. Briefly, each cohort was normalised separately using the Functional Normalization algorithm (73) implemented in meffil. Outlying methylation values (>10 standard deviations from the probe mean) were removed and replaced with the probe mean. Sites measured by sub-optimal probes (67) were removed, as were multi-mapping probes and probes targeting non CpG sites that failed liftover to hg19 (54).

Generating cell count estimates

Cell count estimates were generated using the Houseman algorithm implemented in the R package meffil v0.1.0 (72). Reference panels were as follows: ARIES at birth (74), ARIES at 7 and 15-17 years (75) and both BiB ethnic groups (76).

Adjusting DNAm data for known covariates

DNAm data were adjusted for sex, age (apart from the birth timepoints), sample type if more than one was used (birth and 7 years in ARIES), Beadchip (also referred to as slide) to represent batch effects, and cell count proportion estimates.

Genotype data generation

ARIES participants were genotyped as part of the main ALSPAC study. All ALSPAC child participants were genotyped with the Illumina HumanHap550 quad genome-wide SNP array (Illumina Inc., San Diego, CA) by the Laboratory Corporation of America (LCA, Burlington, NC, USA) and the Wellcome Trust Sanger Institute (WTSI, Cambridge, UK), supported by 23andMe (71); exclusions and imputation procedures are detailed in the supplementary methods. Samples in BiB were genotyped using either the Illumina HumanCoreExome Exome-24 v1.1 microarray, or the Infinium global screen-24+v1.0 array. GenomeStudio 2011.1 was used to pre-process samples; exclusions and imputation procedures are detailed in the supplementary methods.

Correlation of all sites on the 450k array

Please see Figure 8 for an overview of this process. DNAm sites were correlated using the biweight mid-correlation Hardin 2007;Langfelder 2012; Song 2012) (citation, Rpackage WGCNA), a median-based method. As Pearson correlation is mean based it may not be suitable for DNAm data, which may be influenced by genotype and so form clusters. In ARIES, a correlation matrix of 394,842 x 394,842 yields 77,949,905,061 unique correlations when we remove the diagonal of the matrix (in BiB there were 68,374,355,910 unique correlations). Size limits in R mean it is not possible to create

a single matrix containing all pairwise correlations. As a solution, the DNAm sites were split into blocks of 25,000, and all blocks were correlated against each other. To assess the features of correlating pairs, the correlations were then split by value, from -1 to 1, in increments of 0.1. This enabled analysis of the features of all correlations of different strengths (ie, do high correlations differ from low correlations?). The process is summarised in Figure 11, with reproducible code found under the GitHub link.

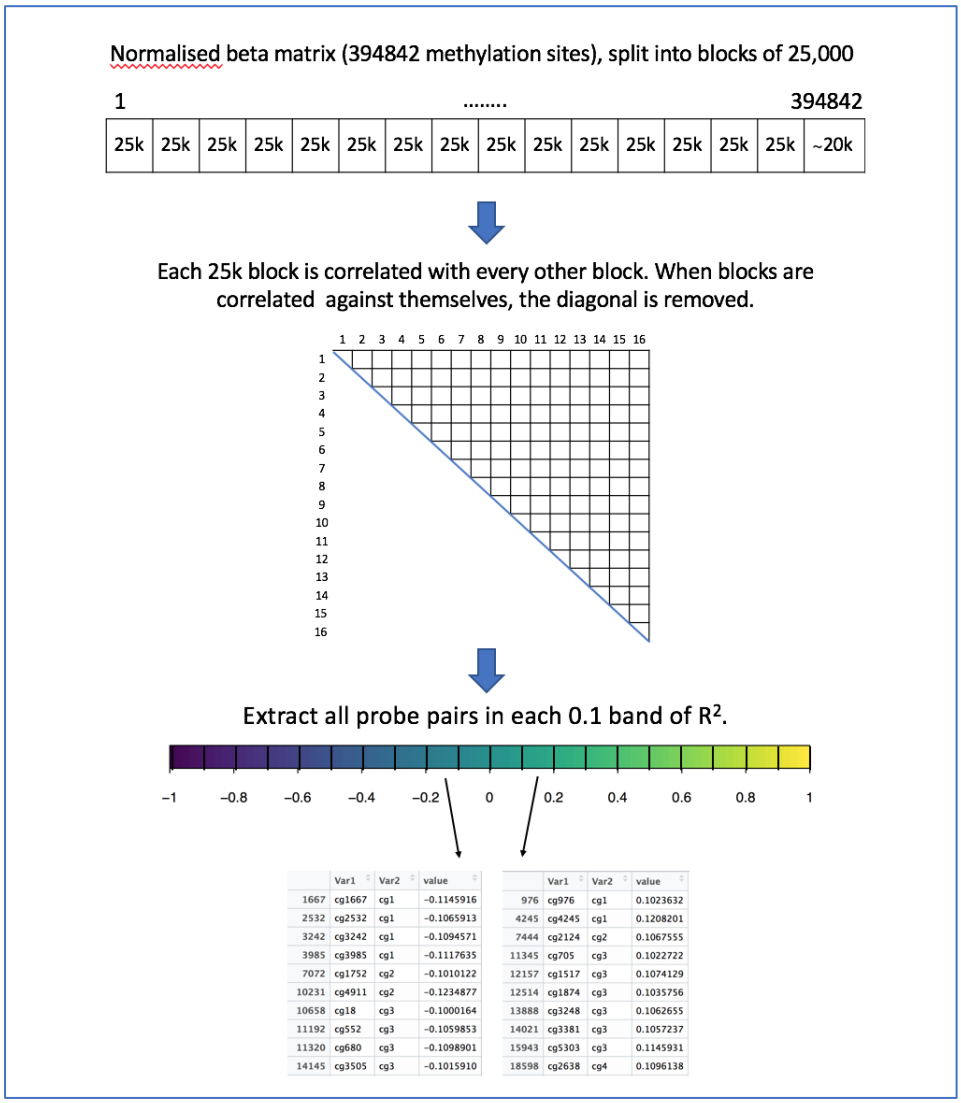

Figure 11: Overview of creating the correlation matrix and extracting pairwise correlations (numbers are representative of the correlation matrix in ARIES).

Creating *cis* decay plots

To assess *cis* correlation structure, decay plots of cis correlations were created across all chromosomes, and for each chromosome separately. This is based on the method used in (Saffari et al., 2018). All correlations where the DNAm sites on the same chromosome were within 10kb were extracted from the data. Positive and negative correlations were then separated so they could be plotted separately, as we hypothesised that negative correlations may not have the same structure as positive *cis* correlations. The pairwise correlations were then binned, specifying at least 1000 pairwise correlations per bin to display all chromosomes, and 100 to display individual chromosomes. Once the correlations were grouped into bins, the mean pairwise correlation for each bin, the standard deviation of the correlation values in each bin, and the median pairwise distance between the correlating pairs of DNAm sites per bin, were calculated. These values were used to create the *cis* decay plots, using a line plot and error bars.

Mean difference plots

Mean difference plots were used to assess the preservation of high correlations between ARIES and BiB, and between the two ethnic groups in BiB. For each pair of DNAm sites, the mean of the correlation of the two groups is plotted against the difference in correlation between the two groups. 95% confidence intervals are calculated using the difference, illustrating whether the correlation differs more between the two groups than expected by chance (77).

Heritability and environmental contributions to co-methylation

To assess the impact of heritability on DNAm correlations, the estimates of heritability and environmental influences on DNAm created by (1) were used to estimate the proportion of sites in each correlation band that were influenced by genetic, unique environmental, and shared environmental factors. The contribution of these influences were assembled for a unique list of all DNAm sites which featured in each correlation range (-1 to 1, in increments of 0.1), and the density of each influence for each correlation range was plotted using a ridgeline plot.

Association of mQTLs with strength of co-methylation

We identified whether, for each correlating pair, neither, one or both of the DNAm sites were associated with an mQTL. For each correlating pair, each DNAm site was assigned 0 if it was not associated with any SNPs in the GoDMC dataset, and a 1 if there were one or more SNP associations. The value was summed for the two sites in a correlating pair, which resulted in 0 if neither DNAm site was associated with a SNP, 1 if only one of the DNAm sites was associated with a SNP, and 2 if both DNAm sites were associated with a SNP. This was done separately for *cis* and *trans* correlations, and totalled over all pairs in each correlation range, to illustrate the distribution of mQTLs across values of correlation. Please note that this does not identify whether both DNAm sites are associated with the same mQTL.

Enrichment analyses of strongly co-methylated sites

To identify whether DNAm sites which form strong correlations overlap with genomic sites of interest, we used the locus overlap package LOLA version 1.10.0 (78). LOLA assesses enrichment based on genomic regions rather than genes. A list of unique DNAm sites correlating >0.9 formed the test dataset; all sites in the analysis formed the background. Because binding of transcription factors is enriched in GC-rich areas of the genome, the content of the background set was reduced and matched to the GC content of the test set using frequency quantiles. Genomic locations of the DNAm sites were taken from the IlluminaHumanMethylation450kanno.ilmn12.hg19 R package version 0.6.0 (79). Start and end sites were computed as -500bp and +500bp from the DNAm site position, respectively. A radius of 1kb was thought to be appropriate overlap because DNAm sites within 1-2kb are highly correlated. We used region sets created by the LOLA team (available through http://lolaweb.databio.org) - the ENCODE transcription factor binding sites (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeAwgTfbsUniform/) , chromHMM imputed 25 chromatin states from Roadmap Epigenomics (https://egg2.wustl.edu/roadmap/data/byFileType/chromhmmSegmentations/ChmmModels/imput

ed12marks/jointModel/final/) (17, 80), and gene annotations from https://zwdzwd.github.io/InfiniumAnnotation.

Assessing *trans* correlations for chromatin contacts

Previous work (16) has illustrated that inter-chromosomal regions which connect have correlated DNAm states. To test whether highly correlating regions are enriched for chromatin contacts, we adapted the GoDMC analysis pipeline (https://github.com/MRCIEU/godmc_phase2_analysis/tree/master/13_hi-c) to test for chromatin contact enrichment, using the chromatin contacts map produced by (81). For each of the pairwise contacts in the (81) dataset the pipeline creates a 1kb region for the two contacting areas of the genome. These are split into files containing all interactions for all possible pairs of chromosomes (for example all contacts between chromosome 1 and chromosome 18). This resulted in 231 files containing all inter- chromosomal contact on the autosomes. Next, the inter-chromosomal correlations r>0.9 in the relevant dataset (e.g. ARIES 7 year olds) had a 500bp region defined either side of the DNAm sites. The pipeline then identified which correlating pairs overlapped with the contact regions (81).

To ascertain whether there were more contacts in our data than expected by chance the pipeline creates a permuted dataset, where from the original data the second DNAm site in the highly correlating pair is replaced randomly with another in the dataset. Broken pairs that match a pair from the original dataset are removed, as are duplicates to avoid double counting. Then overlaps with the (81) data were calculated for the permuted data. This process was repeated 1,000 times, the permuted datasets were merged together, and a distribution of overlap counts was created for the permuted data. This distribution was used to create a p value for the overlap of the permuted distribution chromatin contact overlaps with the number of overlaps in the real data. The permuted distribution of overlaps and the real number of overlaps were plotted using ggplot2 version 3.1.0.

Declarations

Ethics approval and consent to participate

Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees. http://www.bristol.ac.uk/alspac/researchers/research-ethics/ under proposal number B2808. Consent for biological samples has been collected in accordance with the Human Tissue Act (2004). 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.

Consent for publication

Not applicable

Availability of data and materials

The data that supports the findings of this study are available from ALSPAC but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. ALSPAC fully supports Wellcome and the RCUK policies on open access. The process for obtaining access to data is described on the study website: http://www.bristol.ac.uk/alspac/researchers/data-access/. The ALSPAC study website contains details of all the data that are available through a fully searchable data dictionary: http://www.bristol.ac.uk/alspac/researchers/our-data/.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

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. 450K DNAm array data and part of the genotype data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Born in Bradford is only possible because of the enthusiasm and commitment of the Children and Parents in BiB. We are grateful to all the participants, practitioners and researchers who have made Born in Bradford happen. 450K DNAm and genotype array data was generated in the Bristol Bioresource Laboratory Illumina Facility, University of Bristol.

Funding

SHW was funded by a Wellcome Trust PhD studentship (ref 109103/Z/15/Z). TRG and JLM are supported by the UK Medical Research Council Integrative Epidemiology Unit at the University of Bristol (MC_UU_00011/4, MC_00011/5). The funders had no role in the study design, analysis, interpretation of data, or in writing the manuscript. This research was funded in whole, or in part, by the Wellcome Trust [ref 109103/Z/15/Z]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

The UK Medical Research Council and Wellcome (Grant ref: 217065/Z/19/Z) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors and SHW will serve as guarantor for the contents of this paper. A comprehensive list of grants funding is available on the ALSPAC website (http://www.bristol.ac.uk/alspac/external/documents/grant-acknowledgements.pdf). This research was specifically funded by grants enabling DNA methylation data generation: ARIES was funded by the BBSRC (BBI025751/1 and BB/I025263/1), with supplementary funding obtained from the MRC (MC_UU_12013/1 & MC_UU_12013/2 & MC_UU_12013/8), National Institute of Child and Human Development (R01HD068437), NIH (5RO1AI121226-02) and CONTAMED EU collaborative Project (212502). ALSPAC GWAS data was generated by Sample Logistics and Genotyping Facilities at Wellcome Sanger Institute and LabCorp (Laboratory Corporation of America) using support from 23andMe.

BiB receives core funding from the Wellcome Trust (WT101597MA), the British Heart Foundation (CS/16/4/32482), a joint grant from the UK Medical Research Council (MRC) and UK Economic and Social Science Research Council (ESRC) (MR/N024397/1) and the National Institute for Health Research (NIHR) under its Collaboration for Applied Health Research and Care (CLAHRC) for Yorkshire and Humber and Clinical Research Network.

Author contributions

Designed individual studies and contributed data: TRG, JLM, DAL

Designed and managed the study: JLM, NJT, TRG

Designed the analyses: SHW, JLM, TRG, NJT, MS, GH, KB

Conducted analyses: SHW

Critically reviewed and revised the analyses: SHW, JLM, TRG, NJT, MS, GH, KB, DAL

SHW wrote the manuscript; all authors reviewed and revised the manuscript

References

1. Hannon E, Knox O, Sugden K, Burrage J, Wong CCY, Belsky DW, et al. Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. PLoS Genet. 2018;14(8):e1007544.
2. Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. Genome Med. 2014;6(7):60.
3. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines (vol 12, pg R10, 2011). Genome Biol. 2011;12(6).
4. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife. 2013;2:e00523.
5. Saunderson EA, Stepper P, Gomm JJ, Hoa L, Morgan A, Allen MD, et al. Hit-and-run epigenetic editing prevents senescence entry in primary breast cells from healthy donors. Nat Commun. 2017;8(1):1450.
6. Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, et al. Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat Biotechnol. 2013;31(12):1137-42.
7. Bonder MJ, Luijk R, Zhernakova DV, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet. 2017;49(1):131-8.
8. Razin A, Riggs AD. DNA methylation and gene function. Science. 1980;210(4470):604-10.

9. Prendergast GC, Ziff EB. Methylation-sensitive sequence-specific DNA binding by the c-Myc basic region. Science. 1991;251(4990):186-9.
10. Kribelbauer JF, Lu XJ, Rohs R, Mann RS, Bussemaker HJ. Toward a Mechanistic Understanding of DNA Methylation Readout by Transcription Factors. J Mol Biol. 2019.
11. Domcke S, Bardet AF, Adrian Ginno P, Hartl D, Burger L, Schubeler D. Competition between DNA methylation and transcription factors determines binding of NRF1. Nature. 2015;528(7583):575-9.
12. Yin Y, Morgunova E, Jolma A, Kaasinen E, Sahu B, Khund-Sayeed S, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science. 2017;356(6337).
13. Song G, Wang G, Luo X, Cheng Y, Song Q, Wan J, et al. An all-to-all approach to the identification of sequence-specific readers for epigenetic DNA modifications on cytosine. Nat Commun. 2021;12(1):795.
14. Ciccone DN, Su H, Hevi S, Gay F, Lei H, Bajko J, et al. KDM1B is a histone H3K4 demethylase required to establish maternal genomic imprints. Nature. 2009;461(7262):415-8.
15. Smith ZD, Meissner A. DNA methylation: roles in mammalian development. Nat Rev Genet. 2013;14(3):204-20.
16. Li G, Liu Y, Zhang Y, Kubo N, Yu M, Fang R, et al. Joint profiling of DNA methylation and chromatin architecture in single cells. Nat Methods. 2019;16(10):991-3.
17. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
18. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, 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). Hum Mol Genet. 2015;24(8):2201-17.
19. Smith JA, Zhao W, Wang X, Ratliff SM, Mukherjee B, Kardia SLR, et al. Neighborhood characteristics influence DNA methylation of genes involved in stress response and inflammation: The Multi-Ethnic Study of Atherosclerosis. Epigenetics. 2017;12(8):662-73.
20. de Vocht F, Suderman M, Ruano-Ravina A, Thomas R, Wakeford R, Relton C, et al. Residential exposure to radon and DNA methylation across the lifecourse: an exploratory study in the ALSPAC birth cohort. Wellcome Open Res. 2019;4:3.
21. Liu J, Carnero-Montoro E, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nature Communications. 2019;10.
22. Xu CJ, Soderhall C, Bustamante M, Baiz N, Gruzieva O, Gehring U, et al. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. Lancet Respir Med. 2018;6(5):379-88.
23. Battram T, Yousefi P, Crawford G, Prince C, Babei MS, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. OSF Preprints. 2021.
24. Langfelder P, Cantle JP, Chatzopoulou D, Wang N, Gao F, Al-Ramahi I, et al. Integrated genomics and proteomics define huntingtin CAG length-dependent networks in mice. Nat Neurosci. 2016;19(4):623-33.
25. Pidsley R, Viana J, Hannon E, Spiers H, Troakes C, Al-Saraj S, et al. Methylomic profiling of human brain tissue supports a neurodevelopmental origin for schizophrenia. Genome Biol. 2014;15(10):483.
26. Wong CCY, Smith RG, Hannon E, Ramaswami G, Parikshak NN, Assary E, et al. Genome-wide DNA methylation profiling identifies convergent molecular signatures associated with idiopathic and syndromic autism in post-mortem human brain tissue. Hum Mol Genet. 2019;28(13):2201-11.
27. Swarup V, Hinz FI, Rexach JE, Noguchi K, Toyoshiba H, Oda A, et al. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. Nat Med. 2019;25(1):152-+.
28. Eckhardt F, Lewin J, Cortese R, Rakyan VK, Attwood J, Burger M, et al. DNA methylation profiling of human chromosomes 6, 20 and 22. Nat Genet. 2006;38(12):1378-85.

29. Kuan PF, Chiang DY. Integrating prior knowledge in multiple testing under dependence with applications to detecting differential DNA methylation. Biometrics. 2012;68(3):774-83.
30. Liu Y, Li X, Aryee MJ, Ekstrom TJ, Padyukov L, Klareskog L, et al. GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease. Am J Hum Genet. 2014;94(4):485-95.
31. Ong ML, Holbrook JD. Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways. Aging Cell. 2014;13(1):142-55.
32. Saffari A, Silver MJ, Zavattari P, Moi L, Columbano A, Meaburn EL, et al. Estimation of a significance threshold for epigenome-wide association studies. Genet Epidemiol. 2018;42(1):20-33.
33. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20(7):883-9.
34. Zhang W, Spector TD, Deloukas P, Bell JT, Engelhardt BE. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. Genome Biol. 2015;16:14.
35. Garg P, Joshi RS, Watson C, Sharp AJ. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. 2018;14(10):e1007707.
36. Wu Y, Qi T, Wang H, Zhang F, Zheng Z, Phillips-Cremins JE, et al. Promoter-anchored chromatin interactions predicted from genetic analysis of epigenomic data. Nat Commun. 2020;11(1):2061.
37. Shah S, McRae AF, Marioni RE, Harris SE, Gibson J, Henders AK, et al. Genetic and environmental exposures constrain epigenetic drift over the human life course. Genome Res. 2014;24(11):1725-33.
38. van Dongen J, Nivard MG, Willemsen G, Hottenga JJ, Helmer Q, Dolan CV, et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun. 2016;7:11115.
39. Sugden K, Hannon EJ, Arseneault L, Belsky DW, Corcoran DL, Fisher HL, et al. Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. Patterns (N Y). 2020;1(2):100014.
40. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49(2):359-67.
41. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10).
42. Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;22(R1):R7-R15.
43. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8(5):599-618.
44. Houtepen LC, Hardy R, Maddock J, Kuh D, Anderson EL, Relton CL, et al. Childhood adversity and DNA methylation in two population-based cohorts. Transl Psychiatry. 2018;8(1):266.
45. Dunn EC, Soare TW, Zhu Y, Simpkin AJ, Suderman MJ, Klengel T, et al. Sensitive Periods for the Effect of Childhood Adversity on DNA Methylation: Results From a Prospective, Longitudinal Study. Biol Psychiatry. 2019;85(10):838-49.
46. Barcelona de Mendoza V, Huang Y, Crusto CA, Sun YV, Taylor JY. Perceived Racial Discrimination and DNA Methylation Among African American Women in the InterGEN Study. Biol Res Nurs. 2018;20(2):145-52.
47. Needham BL, Smith JA, Zhao W, Wang X, Mukherjee B, Kardia SL, et al. Life course socioeconomic status and DNA methylation in genes related to stress reactivity and inflammation: The multi-ethnic study of atherosclerosis. Epigenetics. 2015;10(10):958-69.
48. Dai L, Mehta A, Mordukhovich I, Just AC, Shen J, Hou L, et al. Differential DNA methylation and PM2.5 species in a 450K epigenome-wide association study. Epigenetics. 2017;12(2):139-48.

49. Baccarelli A, Wright RO, Bollati V, Tarantini L, Litonjua AA, Suh HH, et al. Rapid DNA methylation changes after exposure to traffic particles. Am J Respir Crit Care Med. 2009;179(7):572-8.
50. Tang H. Confronting ethnicity-specific disease risk. Nat Genet. 2006;38(1):13-5.
51. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288-95.
52. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6(6):692-702.
53. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
54. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenomic insights from an atlas of genetic effects on DNA methylation. medRxiv. 2020:2020.09.01.20180406.
55. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol. 2016;17:61.
56. Lienert F, Wirbelauer C, Som I, Dean A, Mohn F, Schubeler D. Identification of genetic elements that autonomously determine DNA methylation states. Nat Genet. 2011;43(11):1091-7.
57. Barabasi AL, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509-12.
58. Barabasi AL. Scale-free networks: a decade and beyond. Science. 2009;325(5939):412-3.
59. Broido AD, Clauset A. Scale-free networks are rare. Nat Commun. 2019;10(1):1017.
60. Turowski TW, Tollervey D. Transcription by RNA polymerase III: insights into mechanism and regulation. Biochem Soc Trans. 2016;44(5):1367-75.
61. Vannini A, Ringel R, Kusser AG, Berninghausen O, Kassavetis GA, Cramer P. Molecular basis of RNA polymerase III transcription repression by Maf1. Cell. 2010;143(1):59-70.
62. Abascal-Palacios G, Ramsay EP, Beuron F, Morris E, Vannini A. Structural basis of RNA polymerase III transcription initiation. Nature. 2018;553(7688):301-+.
63. Maass PG, Barutcu AR, Weiner CL, Rinn JL. Inter-chromosomal Contact Properties in Live-Cell Imaging and in Hi-C. Mol Cell. 2018;69(6):1039-45 e3.
64. Neph S, Stergachis AB, Reynolds A, Sandstrom R, Borenstein E, Stamatoyannopoulos JA. Circuitry and dynamics of human transcription factor regulatory networks. Cell. 2012;150(6):1274-86.
65. Kim S, Dunham MJ, Shendure J. A combination of transcription factors mediates inducible interchromosomal contacts. Elife. 2019;8.
66. Javierre BM, Burren OS, Wilder SP, Kreuzhuber R, Hill SM, Sewitz S, et al. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. Cell. 2016;167(5):1369-84 e19.
67. Zhou W, Laird PW, Shen H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. Nucleic Acids Res. 2017;45(4):e22.
68. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort Profile: the Born in Bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42(4):978-91.
69. Department for Communities and Local Government. The English Indices of Deprivation 2015: Statistical Release 2015 [Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/465791/English_Indices_of_Deprivation_2015_-_Statistical_Release.pdf.
70. Smith T, Noble M, Noble S, Wright G, McLennan D, Plunkett E. The English indices of deprivation 2015. London: Department for Communities and Local Government. 2015.
71. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, et al. Data Resource Profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). Int J Epidemiol. 2015;44(4):1181-90.

72. Min JL, Hemani G, Davey Smith G, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics. 2018;34(23):3983-9.
73. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 2014;15(12):503.
74. Gervin K, Page CM, Aass HC, Jansen MA, Fjeldstad HE, Andreassen BK, et al. Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. Epigenetics. 2016;11(9):690-8.
75. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7(7):e41361.
76. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, S LM, et al. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics. 2016;11(5):354-62.
77. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
78. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. Bioinformatics. 2016;32(4):587-9.
79. Hansen KD. IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. 0.6.0 ed2016. p. R package.
80. Ernst J, Kellis M. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. Nat Biotechnol. 2015;33(4):364-76.
81. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.