Tissue- and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo

Maria Derakhshan, Noah Kessler, Miho Ishida, Charalambos Demetriou, Nicolas Brucato, Gudrun E Moore, Caroline Fall, Giriraj Chandak, Francois-Xavier Ricaut, Andrew M Prentice, et al.

▶ **To cite this version:**

Maria Derakhshan, Noah Kessler, Miho Ishida, Charalambos Demetriou, Nicolas Brucato, et al.. Tissue- and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo. Nucleic Acids Research, 2022, 50 (12), pp.6735-6752. 10.1093/nar/gkac503 . hal-03753713

HAL Id: hal-03753713
https://cnrs.hal.science/hal-03753713

Submitted on 25 Aug 2022

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

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

Title: Tissue- and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo

Authors: Maria Derakhshan[1], Noah J. Kessler[2], Miho Ishida[3], Charalambos Demetriou[3], Nicolas Brucato[4], Gudrun E. Moore[3], Caroline H.D. Fall[5], Giriraj R. Chandak[6], Francois-Xavier Ricaut[4], Andrew M. Prentice[7], Garrett Hellenthal[8*] and Matt J. Silver[1,7*]

Affiliations:

[1]London School of Hygiene and Tropical Medicine, UK.

[2]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK.

[3]UCL Great Ormond Street Institute of Child Health.

[4]Laboratoire Évolution and Diversité Biologique (EDB UMR 5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France.

[5]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, United Kingdom

[6]Genomic Research on Complex Diseases (GRC Group), CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India.

[7]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, The Gambia.

[8]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK.

*Corresponding authors

Corresponding authors' emails:

g.hellenthal@ucl.ac.uk

matt.silver@lshtm.ac.uk

Title: Tissue- and ethnicity-independent hypervariable DNA methylation states show evidence of establishment in the early human embryo

Authors: Maria Derakhshan[1], Noah J. Kessler[2], Miho Ishida[3], Charalambos Demetriou[3], Nicolas Brucato[4], Gudrun E. Moore[3], Caroline H.D. Fall[5], Giriraj R. Chandak[6], Francois-Xavier Ricaut[4], Andrew M. Prentice[7], Garrett Hellenthal[8*] and Matt J. Silver[1,7*]

Affiliations:

[1]London School of Hygiene and Tropical Medicine, UK.

[2]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK.

[3]UCL Great Ormond Street Institute of Child Health.

[4]Laboratoire Évolution and Diversité Biologique (EDB UMR 5174), Université de Toulouse Midi-Pyrénées, CNRS, IRD, UPS, Toulouse, France.

[5]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton, United Kingdom

[6]Genomic Research on Complex Diseases (GRC Group), CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India.

[7]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, The Gambia.

[8]UCL Genetics Institute, University College London, Gower Street, London WC1E 6BT, UK.

*Corresponding authors

Corresponding authors' emails:

g.hellenthal@ucl.ac.uk

matt.silver@lshtm.ac.uk

Abstract

We analysed DNA methylation data from 30 datasets comprising 3,474 individuals, 19 tissues and 8 ethnicities at CpGs covered by the Illumina450K array. We identified 4,143 hypervariable CpGs ("hvCpGs") with methylation in the top 5% most variable sites across multiple tissues and ethnicities. hvCpG methylation was influenced but not determined by genetic variation, and was not linked to probe reliability, epigenetic drift, age, sex or cell heterogeneity effects. hvCpG methylation tended to covary across tissues derived from different germ-layers and hvCpGs were enriched for proximity to ERV1 and ERVK retrovirus elements. hvCpGs were also enriched for loci previously associated with periconceptional environment, parent-of-origin-specific methylation, and distinctive methylation signatures in monozygotic twins. Together, these properties position hvCpGs as strong candidates for studying how stochastic and/or environmentally influenced DNA methylation states which are established in the early embryo and maintained stably thereafter can influence life-long health and disease.

Introduction

DNA methylation (DNAm) plays a critical role in mammalian development, underpinning X-chromosome inactivation, genomic imprinting, silencing of repetitive regions and cell differentiation(1). DNAm states that vary between individuals have been a focus of Epigenome-Wide Association Studies (EWAS) due to their potential to drive phenotypic variation(2, 3). Factors influencing interindividual methylation differences include genetic variation(4, 5), cell heterogeneity effects(6, 7), sex(8, 9), age(10, 11), and pre- and post- natal environment(12–14). Growing evidence from studies investigating DNAm patterns in multiple tissues suggests that these factors can have both shared and tissue-specific influences on DNAm variation (12, 15–18).

In this study, we identified and characterised hypervariable CpGs ('hvCpGs') covered on the widely used Illumina HumanMethylation450K (hereafter 'Illumina450K') array(19) that showed high interindividual variation in multiple datasets covering 19 different tissue/cell types and 8 ethnicities spanning a wide range of ages. We reasoned that identified loci would be robust to both tissue-specific drivers of methylation variability such as those mentioned above and dataset-specific technical artefacts(20–23), thereby revealing insights into biological mechanisms influencing methylation variation across system-wide tissues.

Tissue-independent methylation variation has been previously observed at a class of loci at which methylation not only varies between individuals but is also *correlated* across tissues derived from different germ layers within a given individual. Also described as 'systemic inter-individual variation' or SIV, this property is attributed to stochastic methylation establishment in the pre-gastrulation embryo(24–29). Accordingly, SIV CpGs overlap loci showing 'epigenetic supersimilarity' (ESS) indicating establishment before cleavage in monozygotic (MZ) twins(27) and show sensitivity to the periconceptional environment(24, 25, 27, 28, 30). Several SIV loci have been associated with phenotypic traits and disease, including obesity(31), cancer(25, 27), rheumatoid arthritis(32), autism(33), Alzheimer's disease(34), Parkinson's disease(35) and thyroid volume and function related differences in body fat and bone mineral density (36). SIV loci are therefore promising candidates for exploring the developmental origins of disease, with the additional advantage that easily accessible tissues can be used as proxies for pathologically relevant but inaccessible tissues(37).

We investigated whether hvCpGs showed evidence of establishment in the early embryo and sensitivity to the periconceptional environment. We also examined the genomic context of hvCpGs by exploring their association with multi-tissue histone marks, transposable elements and regions of parent-of-origin specific methylation. Finally, we probed putative functional roles of hvCpGs by interrogating EWAS trait associations and by performing gene ontology enrichment analysis.

Our curated set of hvCpGs show methylation variation that is not explained by probe reliability, age, sex, cell heterogeneity, or genetic effects. Instead, hvCpGs show evidence of establishment in the early embryo and correlation across tissues. They therefore serve as a useful resource for studying the influence of early environmental and/or stochastic effects on DNAm in diverse tissues and ethnicities, and for studying the impact of DNAm differences on life-long health and disease.

Materials and Methods

Methylation data used for identifying hvCpGs

Publicly available methylation Beta matrices were downloaded from The Cancer Genome Atlas (TCGA) (https://www.cancer.gov/tcga) and the Gene Expression Omnibus (GEO)(38) (Supplementary Tables 1 and 2). Methylation Beta matrices were analysed instead of .idat files because Beta matrices are readily available in public databases, and because analysis of multiple datasets with different processing pipelines should strengthen the robustness of our findings of shared high methylation variance across datasets. The TCGA database was used a resource for downloading methylation data from a large number of tissues. TCGA methylation data were downloaded using the *TCGAbiolinks (v2.18.0)* R package(39–41), selecting only samples annotated as 'Solid Tissue Normal'. Of the 33 TCGA datasets, 10 were selected for our study as these had methylation data in at least 20 samples. GEO methylation Beta matrices were downloaded using the *GEOquery (v2.58.0)* R package(42) from 11 unique accessions that were selected to expand both the number of tissues and ethnicities used in our study. Where available, detection p-values (measuring signal intensity), and metadata on age, sex, and disease status were also downloaded. We split GEO beta matrices into separate groups based on ethnicity and tissue/cell type and refer to the resulting 17 separated groups as 'datasets'. Non-public datasets internal to this study include IlluminaEPIC(43) array data from whole blood samples from Gambian 8-9-year olds (ISRCTN14266771(44)) and Illumina450K data from Bornean and Kenyan saliva samples (45) (Supplementary Table 3). These datasets were chosen to expand the number of ethnicities considered in this study. For IlluminaEPIC datasets we selected probes covered on the Illumina450K array. In total, we analysed 30 datasets (3 internal, 10 TCGA and 17 GEO) that covered 8 ethnicities and 19 different tissue/cell types (Supplementary Table 4).

Methylation data processing

For each methylation dataset used in our main analysis, we used the *ChAMP (v2.20.1)* R package(46) to remove: i) probes with a detection p-value > 0.01 in > 5% samples (where detection p-values were available), ii) probes mapping to multiple genomic positions(47), iii) probes mapping to the X and Y chromosomes, and iv) single nucleotide polymorphism (SNP)-related probes identified by Zhou *et al*. (48) that contain SNPs (MAF > 1%) that are within 5 bp of the CpG interrogation site and/or SNPs effecting probe hybridisation. Where ethnicity information was available, we removed probes with population-specific SNPs identified by Zhou *et al.* using 1000 Genomes populations (MAF > 1%), otherwise we removed the General Recommended Probes Probes (48). Probes that had a missing value in any of the samples in a specific dataset were removed from that dataset. To reduce technical biases introduced by differing type I and type II probe designs on the Illumina450K and IlluminaEPIC arrays, we applied Beta Mixture Quantile normalisation (BMIQ)(49) to all datasets using the *champ.norm()* function from the *ChAMP* R package. All datasets were adjusted for the first 10 principal components (PCs) of variation to account for methylation variability driven by known and/or

unknown technical artefacts (such as plate and array position) and cell heterogeneity. Methylation values were adjusted for these 10 PCs, age (where available) and sex by taking the residuals from a single linear regression model on methylation M values, where M is defined as $log_2(beta/1\text{-}beta)$ (see Supplementary Tables 1-3 for details of the linear model applied to each dataset). Adjusted M values were transformed back into Beta values by applying the transformation *exp(adjusted_M)/(1+exp(adjusted_M)).* Finally, for each probe, we removed outlier methylation values, defined according to Tukey's outer fences (Q1 – 3*IQR and Q3 + 3*IQR). The hg19 reference genome was used throughout all relevant analyses as the Illumina450K array metadata manifest uses this version.

Identification of hvCpGs

We defined an hvCpG in the following way:

(1) in $\geq$ 65% of datasets in which the CpG is covered (following quality control), it has methylation variance in the top 5% of all (non-removed) CpGs.

(2) is covered in at least 15 of the 30 datasets.

While the (5%, 65%) threshold in (1) is arbitrary, we note that ~80% of the resulting set of hvCpGs were also captured when using a different (20%, 90%) threshold, indicating that these hvCpGs are in the top 20% of variable CpGs in $\geq$ 90 % of datasets they are covered in (Supplementary Fig. 1).

Probe reliability

Technically unreliable probes were identified by examining intra-class correlation coefficients (ICCs) from two studies. The first study compared methylation consistency between the Illumina450K and IlluminaEPIC platforms using 365 blood DNA samples, defining poor quality probes as those with ICC $\leq$ 0.4(23). The second study examined methylation reliability between technical replicates from 265 African American peripheral blood leukocyte samples on the Illumina450K platform, defining poor quality probes as those with ICC $\leq$ 0.37(50). We defined technically unreliable probes as those reported as being poor quality in at least one of these two studies.

Methylation quantitative trait locus (mQTL) analysis

mQTL summary statistics from the Genetics of DNA Methylation Consortium (GoDMC), a meta-GWAS of 36 European blood cohorts (N = 27,750) generated using imputed genotype data (~10 million SNPs) and ~420,000 CpGs (51) were used for this analysis. Significance thresholds of $p < 1x10^{-8}$ and $p < 1x\ 10^{-14}$ were applied for *cis* and *trans* mQTLs respectively(51), giving 271,724 significant SNP-CpG associations comprising 190,102 CpGs and 224,648 SNPs. The variance in DNA methylation explained by a given mQTL was estimated as $2 * \beta * MAF(1\text{-}MAF)$, where β is the effect size and MAF is the minor allele frequency(52). To investigate mQTL effects acting at a given CpG, we first calculated the % variance explained by each associated mQTL before calculating the mean % variance explained across all mQTLs associated with the CpG.

Monozygotic twin discordance

We analysed CpGs identified as being 'equivalently variable' between MZ co-twins and between unrelated individuals ('evCpGs (blood)') by Planterose Jiménez *et al.*(53) using Illumina450K data in whole blood. 154 of these evCpGs replicated in adipose tissue from 97 MZ twin pairs ('evCpGs (blood & adipose)'). evCpGs are candidates for methylation states that are established stochastically after MZ twin splitting and are used in our study to indicate CpGs at which genetic effects do not play a large role in methylation variation.

Control CpG sets

Distribution-matched controls

To ensure that several of our analyses are not biased by distributional properties of hvCpGs such as their high variability and enrichment for intermediate methylation states (Supplementary Fig. 2), we constructed a set of CpGs with similar distribution of methylation Beta values to hvCpGs in the Caucasian blood dataset ('Blood_Cauc', Supplementary Table 1). This dataset was chosen as it has the highest number of post-natal samples and because several downstream analyses leverage published studies that used blood methylation data. For each of the 4,108 hvCpGs covered in the 'Blood_Cauc' dataset, a two-sided Kolmogorov-Smirnov (KS) test (*ks.test()* in R) was used to test for the divergence in methylation Beta distributions between the hvCpG and each technically reliable (see 'Probe reliability', Methods) background CpG, selecting the background CpG with the greatest p-value (requiring a p-value > 0.1). In total, 3,566 hvCpGs were each matched to a control CpG ('distribution-matched controls', Table 1, Supplementary Fig. 3).

mQTL-matched controls

To determine the degree to which hypervariability at hvCpGs is explained by mQTL effects, each hvCpG was matched to a CpG amongst those reported in the GoDMC meta-analysis(51). Controls were selected to have i) the same number of mQTL associations, ii) a similar mean % variance explained by mQTL (across all significant mQTL) and iii) presence in at least as many datasets as the hvCpG (Table 1, Supplementary Fig. 4).

Identification of hvCpG clusters

hvCpG clusters were identified by considering the decay of methylation correlation with distance at hvCpGs. To do this, we calculated the average pairwise Spearman correlation (ρ) across hvCpG pairs with inter-CpG distance falling within 100 bp bins, for datasets with at least 100 samples (Supplementary Fig. 5A). The distance threshold for defining hvCpG clusters was chosen to be 4,000 bp as this is approximately the point at which pairwise correlations levelled out (Supplementary Fig. 5A). In total, 2,219 (54%) hvCpGs fell into 716 clusters comprising at least 2 CpGs, with the remaining 1,924 (46%) hvCpGs falling outside of these clusters

(Supplementary Fig. 5B). In 563 (79%) of these clusters, the average Spearman correlation (ρ) across hvCpG pairs was > 0.5 (Supplementary Fig. 5C).

'De-clustering' of hvCpGs

To account for the possibility that our analyses may be biased by the non-random distribution and inter-dependence of hvCpGs in CpG clusters, we generated a de-clustered set of hvCpGs in which no CpG was within 4 kb of another CpG. 2,640 de-clustered hvCpGs were generated by randomly selecting one CpG from each of the clusters and then including all 'singleton' CpGs falling outside of clusters.

Age stability

To examine temporal stability of hvCpGs we used published intra-class correlation coefficients (ICCs) for probes on the Illumina450K array determined using white blood cell samples taken ~6 years apart(54). The ICC scores compare within-sample variability (across the two time-points) to between-sample variability, with ICC $\geq$ 0.5 defined as temporally stable by Flanagan *et al.* (54). To account for the possibility that high ICC scores might be driven by the high variability of hvCpGs, we compared ICC scores at hvCpGs to those at CpGs with similar methylation Beta distributions to hvCpGs at the first time point (Supplementary Fig. 6A). These CpGs were matched to each hvCpG using the same Kolmogorov-Smirnov method detailed in 'Distribution-matched controls' but using publicly available Flanagan *et al*. methylation data (GSE61151) instead the 'Blood_Cauc' dataset (54).

While we regressed out the effect of age in those datasets where this covariate was available, we checked for potential residual age effects on both methylation mean and variance ('epigenetic drift') by a) performing sub-analyses of infant and cord blood datasets (Supplementary Table 10); and b) determining the proportion of hvCpGs that overlap a published set of 6,108 CpGs identified using whole blood Illumina450K data from 3,295 individuals aged 18 to 88 years that show an increased methylation variability with age of more than 5% every 10 years (11) (Supplementary Fig. 6B).

Sex effects

We regressed out the effect of sex in those datasets where this was available. However, since sex is an important potential driver of inter-individual methylation differences, we further examined the potential for hvCpG methylation to be driven by sex-specific effects in 8 datasets that had an approximately equal number of male and female samples and a sample size > 80 (Blood_Japan, Blood_Mexican, Blood_Gamb, CD4+_Estonian, CD8+_Estonian, Saliva_Cauc, Buccals_Sing_9mo, Buccals_Cauc). We split each dataset by sex to generate 8 'male-only' and 'female-only' sub-datasets. We then calculated the proportion of hvCpGs that had methylation variance in the top 5% in each of these sub-datasets.

Published CpG sets used to investigate early embryo establishment

We used the following publicly available data to examine evidence that methylation states at hvCpGs are established in the early embryo. See Table 2 for a summary of these datasets. We note that not all hvCpGs may have been covered in the array background of all of these studies.

Systemic Interindividual Variation ('SIV') CpGs

SIV-CpGs were collated from four published datasets that used either whole genome bisulfite sequencing (WGBS) or Illumina450K data from multiple tissues derived from different germ layers to identify CpGs displaying high interindividual variation and low intra-individual (cross-tissue) variation. These properties are suggestive of variable methylation establishment before germ layer differentiation(26–29). Further details on the four SIV screens used in this study are given in Supplementary Table 6.

Epigenetic Supersimilarity ('ESS') CpGs

Epigenetic supersimilarity (ESS) loci were identified by van Baak *et al.* (27) using Illumina450K data from adipose tissue from 97 MZ and 162 dizygotic (DZ) twin pairs (55). In that study, 1,580 ESS sites were identified within the top decile of methylation variance, with an interindividual methylation range > 0.4 and greater-than-expected concordance in MZ twins vs DZ twins. This supersimilarity amongst MZ twins is attributed to methylation establishment before MZ twin splitting.

MZ twinning CpGs

Van Dongen *et al.* (56) performed an epigenome-wide association analysis on each of 6 cohorts with methylation data from both MZ and DZ twins (5 blood and 1 buccal) to identify probes differentially methylated between MZ twins and DZ (dizygotic) twins. A meta-analysis was then performed using the blood datasets to identify 834 Bonferroni-significant differentially methylated CpGs, which we refer to as 'MZ twinning CpGs'.

Season of conception ('SoC') CpGs

Silver *et al.* (57) used Illumina450K data to identify 259 CpGs associated with season-of-conception ('SoC') in Gambian 2-year olds, each of which showed a minimum methylation difference of 4% between individuals conceived in the peaks of the Gambian rainy and dry seasons.

Transposable elements and telomeres

Locations of ERV1 and ERVK transposable elements determined by RepeatMasker were downloaded from the UCSC annotations repository as previously described(28). Telomere coordinates were downloaded from the UCSC hg19 annotations repository (http://genome.ucsc.edu).

Imprinted genes, parent-of-origin-specific methylation (PofOm)

Imprinted genes classified as 'predicted' or 'known' were downloaded from https://www.geneimprint.com. Parent-of-origin-specific CpGs were identified by Zink *et al.* (58) using WGBS data from peripheral blood from Icelandic individuals.

SIV power calculation

To assess power to detect SIV in previous screens with small numbers of samples, we analysed the 4-individual multi-tissue dataset used by van Baak *et al*. (27, 59). We downloaded this dataset from GEO (GSE50192), selecting the same tissues (gall bladder, abdominal aorta sciatic nerve) used by van Baak *et al*.(27). For each of the 1,042 SIV-CpGs reported by van Baak *et al*., we generated methylation values for three tissues for each simulated individual by randomly sampling from a 3-dimensional multivariate normal distribution, with mean equal to the mean of each tissue's sampled methylation values at the CpG, and standard deviation specified by a 3x3 cross-tissue co-variance matrix of the sampled methylation values at the CpG. For each SIV-CpG, we sampled four simulated individuals and determined if this random sample met the SIV definition specified by van Baak *et al*.(27), repeating this process 1000 times to give a power estimate (Supplementary Fig. 7).

Processing and analysis of fetal multi-tissue dataset

The unpublished fetal multi-tissue dataset comprised 60 samples, corresponding to 30 individuals that each have methylation data from two tissues derived from different germ layers (ectoderm: brain, spinal cord, skin; mesoderm: kidney, rib, heart, tongue; endoderm: intestine, gut, lung, liver). These fetal tissues were obtained from the 'Moore Fetal Cohort' from the termination of pregnancies at Queen Charlotte's and Chelsea Hospital (London, UK). Ethical approval for obtaining fetal tissues was granted by the Research Ethics Committee of the Hammersmith, Queen Charlotte's and Chelsea and Acton Hospitals (2001/6028). DNA was extracted from fetal tissues using the AllPrep DNA/RNA/Protein Mini Kit (Qiagen) and bisulfite conversion was carried out using EZ DNA Methylation Kits (Zymo Research). Samples were then processed using the Illumina InfiniumEPIC array. Derived methylation data were imported as *.idat* files into R and analysed using the *meffil* R package (*v 1.1.2*)(60) with default parameters. Briefly, methylation predicted sex was used to remove 2 sex outliers (samples with methylation > 5 SDs from mean). Next, 1 sample was removed for which the predicted median methylation signal was more than 3 SDs from the expected signal, leaving 57 samples. 515 probes with detection-p-value value > 0.1 and 307 probes with bead number < 3 in more than 20% of samples respectively were removed. Array data were then corrected for dye-bias and background effects and functional normalisation was applied, specifying the number of PCs to be 7 (the PC at which the variance explained at control probes levelled out). Next, the *ChAMP (v2.20.1)* R package(46) was used to remove cross-hybridising and multi-mapping probes, probes on XY chromosomes, and SNP-related probes, leaving 746,492 CpGs. We selected the 452,016 probes that overlapped the Illumina450K array and the 27 individuals for which both tissue samples passed quality control. This included 9 individuals with methylation data from endoderm and mesoderm, 10 individuals with methylation data from endoderm and ectoderm and 8

individuals with methylation data from mesoderm and ectoderm (see Supplementary Table 7). Methylation was then adjusted for predicted sex and batch using a linear model. The mean gestational age of these individuals was 12.5 weeks.

For the 9 individuals with available endoderm-mesoderm samples, we calculated the Pearson r between germ layer methylation values for each hvCpG and repeated this for individuals with endoderm-ectoderm and mesoderm-ectoderm samples. The inter-germ layer correlation was then defined as the average Pearson r across these three comparisons. We calculated interindividual variation using the same metric as van Baak *et al*.(27): for each CpG, we took the mean methylation value across the two germ-layer derived tissues for every individual (giving 27 values for each CpG) and defined interindividual variation of the CpG as the range of these means.

Chromatin states at hvCpGs

Chromatin states were predicted by a ChromHMM 15-state model(61) using Chromatin Immunoprecipitation Sequencing (ChIP-Seq) data generated by the Roadmap Epigenomics Consortium (62). These data were downloaded for H1 ESCs (E003), fetal brain (E071), fetal muscle (E090), fetal small intestine (E085), foreskin fibroblasts (E055), adipose (E063) and primary mononuclear cells (E062) from the Washington University Roadmap repository. Chromatin states were collapsed into 8 states for clarity (Supplementary Table 8).

EWAS trait associations at hvCpGs

hvCpG trait associations were determined using the EWAS catalogue (http://ewascatalog.org/), which details significant results (p-value < 1 x 10^{-4}) from published EWAS studies. Considering only those traits for which at least 1% of hvCpGs overlapped associated CpGs (highlighted in green in Supplementary Table 9), we first extracted the array background CpGs overlapping the 'Blood_Cauc' dataset that were associated with each trait. We then calculated enrichment odds ratios of hvCpGs relative to blood distribution-matched controls (Table 1) and determined the significance of the enrichment using Fisher's Exact Tests.

GTEx transcription levels

The median gene transcription levels for 54 tissues were downloaded from the GTEx portal (https://gtexportal.org/home/datasets). Transcription levels were examined at 416 out of the 425 genes that were annotated to an hvCpG cluster in the Illumina450K manifest.

Gene ontology term enrichment analysis

Gene Ontology (GO) term enrichment analysis was performed using the *missMethyl* R package (*v1.24.0*)(63) using the *gometh()* function, setting arguments sig.cpg = hvCpGs, all.cpg = array.background, sig.genes = T, collection = "GO", array.type = "450K" and prior.prob = T to adjust for variation in the number of 450K probes mapping to each gene.

Bootstrapped confidence intervals

All bootstrapped 95% confidence intervals were calculated over 1,000 bootstrap samples.

Results

Identification of hypervariable CpGs

We analysed methylation data from 3,474 individuals across 30 datasets (28 Illumina450K and 2 EPIC array) comprising 19 unique tissue/cell types and 8 ethnicities covering a range of ages (Supplementary Tables 1-4). We focussed on CpGs covered by the Illumina450K array and began by excluding probes with poor detection p-values, cross-hybridising probes, probes on the X and Y chromosomes and probes associated with known SNPs (see Methods for details).

We aimed to identify CpGs with consistently high interindividual variation in methylation across diverse datasets, so minimising the effects of dataset-specific drivers of variability including those related to different normalisation methods and processing pipelines. Reasoning that removal of unmeasured technical, batch and cell heterogeneity effects would maximise power to detect true variable methylation states, we adjusted all methylation values for the first ten principal components (PCs) of methylation variation, and additionally adjusted for sex (in datasets with both sexes) and age (where available).

Our strategy for identifying tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs') is summarised in Fig. 1 and detailed in 'Methods'. We defined hvCpGs as CpGs with methylation Beta variance in the top 5% of all CpGs in at least 65% of datasets in which the CpG was covered (Table 1), yielding 4,330 hvCpGs. Note that no CpGs are expected to meet these criteria if the top 5% most variable CpGs in each dataset are entirely independent of those in the others. These thresholds are arbitrary but were chosen in order to select CpGs that met our required criteria of being highly variable in a large number of tissues (median = 13, IQR = [10,15]) and ethnicities (median = 7, IQR = [[6,7]) (Fig 1B). Further, we note that ~80% of identified hvCpGs were within the top 20% of variable CpGs in at least 90% of datasets (Supplementary Fig. 1), meaning that the majority of hvCpGs are within the top 20% of variable loci in almost all covered datasets.

We next compared the set of 4,330 hvCpGs with an alternative set obtained using the same method but without prior adjustment of each dataset for the first ten PCs. This alternative set contained only 1,302 CpGs, which confirmed our intuition that PC adjustment maximises power to identify true dataset-independent hypervariability by removing unwanted technical variation (Supplementary Fig. 8). Finally, we used reported measures of methylation variability among technical replicates (23, 50) to remove 187 technically unreliable probes (see Methods), leaving a final set of 4,143 hvCpGs (Table 1; Supplementary Table 5).

hvCpGs are enriched for intermediate methylation values in all datasets compared to the array background (Supplementary Fig. 2; see Table 1 for definition of array background) and are distributed throughout the genome (Supplementary Fig. 9A-B), with 2,219 (54%) falling within 716 'clusters' containing two or more hvCpGs separated by < 4 kb (Supplementary Fig. 5B). To account for the possibility that our downstream

analyses may be biased by these distributional properties, we generated a set of controls that were distribution-matched in a whole blood dataset (Supplementary Fig. 3) and a set of 'de-clustered hvCpGs' (Table 1, 'Methods').

hvCpG variability is not driven by age, sex, or cell heterogeneity

Evidence from multiple studies suggests that methylation variability can increase with age (termed epigenetic drift)(11, 64), raising the possibility that cross-dataset hypervariability of hvCpGs is driven in part by a large proportion of adult/elderly samples. However, 3,815 (92%) out of 4,122 hvCpGs with methylation measured in cord blood and/or buccal samples from infants showed methylation variance within the top 5% of CpGs in those datasets (Supplementary Table 10), suggesting that high variability at hvCpGs arises in early life. We further probed age stability of hvCpGs by leveraging two studies of age effects in blood. The first study reported methylation consistency in individuals sampled at two time points six years apart using intraclass correlation coefficients (ICCs)(54). Because ICCs increase with CpG variability, we compared temporal stability of hvCpGs to controls with similar methylation Beta distributions selected at the first time point ('Methods'). The temporal stability of hvCpGs was significantly greater than that of controls (Wilcox paired signed-rank test p-value < 5.7×10^{-81}), with 95% of hvCpGs considered temporally stable versus 89% of controls (Supplementary Fig. 6A). The second measured epigenetic drift in a cross-sectional study of 3,295 whole blood samples from individuals aged 18 to 88(11). Only 7% of hvCpGs overlapped CpGs that show increased methylation variability with age, compared to 16.5% of blood distribution-matched controls (Supplementary Fig. 6B). This suggests that that the majority of hvCpGs are stable over a broad time period in whole blood and further supports the notion that hypervariability of hvCpGs in multiple datasets is not an artefact of epigenetic drift effects.

Methylation values were pre-adjusted for the first ten PCs and for sex in all datasets where sex was available as a covariate (24 out of 30 datasets). We investigated the potential for unaccounted-for sex effects to drive methylation variance at hvCpGs by constructing male-only and female-only datasets ('Methods'). 100% of hvCpGs were in the top 5% of CpGs by methylation variance in at least one of the 'male-only' and 'female-only' datasets analysed. Similarly, 3,548 (96%) of the 3,678 hvCpGs covered in purified CD4+ and CD8+ datasets had methylation variance among the top 5% in at least one dataset (Supplementary Table 10), suggesting that methylation variation at hvCpGs was not driven by unaccounted-for cell heterogeneity effects amongst the heterogeneous tissue types studied.

Together, these data strongly suggest that variability at hvCpGs is not driven by sex, age, or cell heterogeneity effects.

Hypervariability is not driven by genetic variants

Genetic variation is an important driver of interindividual methylation differences (4, 5). There is evidence that mQTLs can be shared across different tissues(15, 16, 65, 66) and ethnic groups (5), raising the possibility that 'universal' (multi-tissue and multi-ethnic) mQTLs might drive cross-dataset variability at hvCpGs. We therefore investigated the potential influence of methylation quantitative trait loci (mQTL) on methylation variability at hvCpGs by leveraging a recently published large meta-GWAS (36 cohorts, n = 27,750 individuals) that identified common genetic variants associated with methylation in blood from Europeans(51), reasoning that by definition 'universal' mQTLs would be included in this meta-analysis.

We considered multiple methylation variance thresholds (5%, 10% and 20%) and observed a positive relationship between hypervariability and both the probability of a significant mQTL association and the mean mQTL effect size (Fig. 2A). Amongst the set of 4,143 hvCpGs, there were 6,985 *cis* mQTL (covering 3,635 hvCpGs and 6,417 SNPs) and 971 *trans* mQTL (covering 713 hvCpGs and 753 SNPs). Overall, 3,722 (90%) hvCpGs were reported to be associated with at least one (*cis* or *trans*) mQTL. The median of the mean % variance explained by mQTLs was 4% (Fig. 2B), suggesting that additive genetic effects explain a small to moderate proportion of methylation variability at the majority of these hypervariable loci in blood. Noting that the statistical power to detect mQTL associations will be greater at loci that are inherently variable, we matched hvCpGs to CpGs with the same number of mQTL associations and similar mean % variance explained by mQTL ('mQTL-matched controls', Table 1, Supplementary Fig. 4). hvCpGs showed an average 5-fold increase in methylation variance compared to mQTL-matched controls across datasets (Fig. 2C), further supporting the notion that methylation variation at hvCpGs is not principally driven by universal genetic effects.

To further probe the influence of genetic effects on hvCpG methylation we examined the overlap between hvCpGs and a published set of CpGs that show DNAm variation between monozygotic (MZ) co-twins that is *equivalently variable (ev)* to that between unrelated individuals, suggestive of genetically independent variable methylation establishment after MZ twin splitting(53). In total, hvCpGs comprise 122 (42%) of the 317 evCpGs identified in blood (1.9-fold enrichment relative to distribution-matched controls) and 62% of those that were replicated as evCpGs in adipose tissue (2.8-fold enrichment relative to controls) (Supplementary Table 11), supporting the notion that hvCpGs are likely influenced but not determined by genetic variation in multiple tissues.

hvCpGs show covariation across tissues derived from different germ layers

DNAm states that are variable in different tissues and that are influenced but not determined by genotype may have been established before germ layer separation in early embryonic development and may therefore covary across tissues derived from different germ layers (28). None of the 30 datasets used to identify hvCpGs had multi-tissue data from the same individuals. We therefore examined the overlap between hvCpGs and 3,089 CpGs that show systemic (cross-tissue) interindividual variation (SIV), collated from four published sources(26–29) (Supplementary Table 6). Because both SIV-CpGs and hvCpGs are enriched for intermediate methylation states (28), we used the set of blood distribution-matched controls (Table 1) as a comparator to ensure that our analysis was not biased by this shared property. 24% of hvCpGs overlap a known SIV-CpG, showing a ~5-fold enrichment for SIV-CpGs relative to blood distribution-matched controls (Fig. 3A, Supplementary Table 12, Supplementary Fig. 10A). We note that a further 540 (13%) hvCpGs are within 1 kb of a SIV-CpG, ~5-fold greater than array background CpGs. This suggests that many hvCpGs directly overlap or co-localise with a known SIV-CpG.

The set of all hvCpGs comprises 32.1% of the 3,089 CpGs reported as SIV in any of the four independent studies analysed despite comprising <1% of the 450K array. When considering 'high-confidence' SIV-CpGs reported in at least two or three of the four screens, the proportion identified rises to 76.5% and 95.1% respectively (Fig. 3B). This suggests that our approach of identifying hypervariable loci across multiple datasets may be a more powerful method for identifying putative SIV loci, compared to existing SIV screens that necessarily rely on rare datasets with multi-tissue, multi-germ layer methylation data from small numbers of individuals. To confirm this, we estimated the power to detect SIV using the multi-tissue data from four individuals analysed by van Baak *et al.* (27). Using a permutation framework ('Methods'), we estimated the mean power to detect SIV as 56% (median [IQR] = 0.58 [0.44, 0.72]; Supplementary Fig. 7). As expected, given the small sample size of this multi-tissue dataset, a large proportion of hvCpGs (75%) did not meet the minimum interindividual variation threshold of 0.2 used by van Baak *et al*. to define SIV. On the assumption that hvCpGs are highly enriched for true SIV, this could explain why hvCpGs constitute 61.7% of the van Baak *et al.* SIV-CpGs, while just 13.5% of hvCpGs are identified as SIV-CpGs in the van Baak *et al*. analysis.

To directly test our hypothesis that hvCpGs comprise previously unidentified SIV loci, we analysed a dataset of fetal tissues from 27 individuals, each with methylation data from two tissues derived from different germ layers (see Supplementary Table 7). Inter-germ layer correlations at hvCpGs had a median average Pearson r of 0.42, compared to array background CpGs which had a median average Pearson r of 0.05 (Fig. 3C left). Of the 3,878 hvCpGs covered in this fetal multi-tissue dataset, 1,653 (42%) had an average inter-germ layer Pearson r $\geq$ 0.5. Of these, 58% did not overlap previously identified SIV loci, suggesting that hvCpGs

comprise novel SIV loci. A comparison of the average inter-germ layer correlation at hvCpGs and at previously identified SIV-CpGs showed that hvCpGs and SIV-CpGs had similar inter-germ layer correlations (Fig. 3C right).

hvCpGs are enriched for loci with distinctive methylation patterns in MZ twins

We further investigated evidence for establishment of hvCpG methylation states in the early embryo by testing the overlap between hvCpGs and a published set of 1,217 "epigenetic supersimilarity" (ESS) CpGs overlapping array background. ESS CpGs show high interindividual variation with greater-than-expected methylation concordance between monozygotic co-twins in adipose tissue, suggestive of methylation establishment in the early zygote before MZ cleavage (27). 13% of hvCpGs overlap an ESS CpG, showing a ~9.5-fold enrichment for ESS CpGs relative to distribution-matched controls (Fig. 3A, Supplementary Table 12, Supplementary Fig. 10B).

We next examined the overlap between hvCpGs and a published set of CpGs showing a unique methylation signature in adult tissues from MZ vs DZ twins ('MZ twinning CpGs', Table 2), linked to MZ twin splitting events in early development(56). 7% of hvCpGs overlap an MZ twinning CpG, showing a 3.7-fold enrichment for MZ twinning CpGs compared to distribution-matched controls (Fig. 3A, Supplementary Table 12).

Notably, 54% of ESS and 37% of MZ twinning CpGs overlapping array background are hvCpGs (Fig. 3B).

Reconciling the timing of variable methylation establishment at hvCpGs

The enrichments that we observe for SIV, ESS, evCpGs and MZ twinning CpGs offer a potential insight into the timing of methylation establishment at hvCpGs. In total, 38% of hvCpGs overlap at least one of these CpG sets (Supplementary Fig. 11A) and enrichment is stronger amongst CpGs that show at least two of these properties (Supplementary Fig. 11B). In particular, hvCpGs comprise 78% of SIV-ESS loci and 65% of SIV-MZ twinning loci, suggesting that SIV loci with evidence of establishment in the pre-gastrulation embryo are enriched for hvCpGs.

Variable methylation states identified at evCpGs are thought to originate in embryonic development and/or early post-natal life(53). We note that 41 out of 317 evCpGs overlap SIV and/or MZ twinning CpGs, suggesting that at least a subset may be established in the pre-gastrulation embryo. hvCpGs comprise 67% of evCpGs that overlap SIV-CpGs, and 76% of that overlap MZ twinning CpGs (Supplementary Fig. 11B).

hvCpGs are enriched for parent-of-origin methylation and proximal TEs

In mice, variable methylation states have been associated with the Intracisternal A Particle (IAP) class of endogenous retrovirus(67, 68), with growing evidence that methylation variability may in part be driven by incomplete silencing of IAPs in early development(69, 70). In humans, SIV-CpGs are enriched for proximal endogenous retrovirus elements (ERVs), including the subclasses ERV1 and ERVK(28). This is also the case with hvCpGs: 45% and 7% of hvCpGs are located within 10 kb of an ERV1 and ERVK element respectively, representing a ~1.3-fold and ~1.7-fold enrichment relative to both array background and blood distribution-matched controls (Fig. 3D, Supplementary Fig. 10 C, Supplementary Table 12). Approximately 4.7% of hvCpGs are also located within 1Mb of telomeric regions, showing a 1.8-fold enrichment relative to distribution-matched controls and array background CpGs (Supplementary Table 12).

Maintenance of parent of origin-specific methylation (PofOm) in the pre-implantation embryo is critical for genomic imprinting (71), and several previously identified SIV loci have been found to be associated with imprinted genes and/or PofOm (25, 27, 57). 58 hvCpGs (1.4%) were annotated to 32 imprinted genes (Supplementary Table 13), no more than expected by chance since 1.9% of array background CpGs are annotated to imprinted genes. 10 hvCpGs were annotated to the polymorphically imprinted non-coding RNA *VTRNA2-1,* a well-established SIV locus that is associated with periconceptional environmental exposures(25, 27, 30, 72, 73). Although only a small proportion (2.2 %) of hvCpGs overlap regions of PofOm identified in peripheral blood(58), this overlap represents a 3.5-fold and 11-fold enrichment relative to distribution-matched controls and array background respectively that is maintained after de-clustering (Fig. 3A, Supplementary Fig. 10D, Supplementary Table 12). This overlap constitutes 13% of all PofOm CpGs overlapping array background (Fig. 3B).

hvCpGs show sensitivity to pre-natal environment

Variable methylation states established in early development that are sensitive to environmental perturbation are promising candidates for exploring the developmental origins of health and disease (74–76). We explored whether hvCpGs show sensitivity to pre-natal environment by examining their overlap with loci associated with season of conception ('SoC') in a rural Gambian population exposed to seasonal fluctuations in diet and other factors (77–79). hvCpGs comprise 70 (29%) out of 242 previously identified SoC-CpGs (57) overlapping array background, an approximately 3-fold enrichment relative to distribution-matched controls (Supplementary Table 11).

We next leveraged a recent meta-analysis of 2,365 cord blood samples that modelled genetic (G), genetic by environment (GxE) and additive genetic and environment (G+E) effects at variably methylated probes, where

E represents a range of prenatal exposures including pre-pregnancy BMI, maternal smoking, gestational age, hypertension, anxiety and depression(14). Of the 703 hvCpGs overlapping the neonatal blood variably methylated regions explored in that study, G, GxE, and G+E effects were the 'winning' models for 30%, 30% and 40% of probes respectively, representing an increase in G+E effects compared to array background (Supplementary Fig. 12). This analysis supports our intuition that hvCpGs are influenced but not determined by genetic variation, with pre-natal environment as an additional influencing factor.

Chromatin states at hvCpGs

Compared to array background, hvCpGs are enriched within intergenic regions and CpG island 'shores' but are depleted within gene bodies and regions directly upstream of transcription start sties (Supplementary Fig. 9C). We predicted chromatin states at hvCpGs by examining the overlaps of hvCpGs with histone modifications using the chromHMM 15-state model (61) for seven tissues including embryonic stem cells (H1 ESCs), and fetal and adult tissues(62). Although many hvCpGs were associated with regulatory elements in all tissues, hvCpGs were generally depleted in these regions compared to array background, except within predicted enhancers in H1 ESCs (Supplementary Fig. 13).

Gene expression and ontology analysis

409 hvCpG clusters (corresponding to 1,282 CpGs) are annotated to 425 genes in the Illumina450k manifest. Analysis of GTEx expression data reveals that these are expressed in a diverse range of tissues. (Supplementary Fig, 14). Gene ontology enrichment analysis revealed that hvCpGs were significantly enriched for terms associated with cell-cell adhesion (Fig. 4A), which is largely driven by the colocalization of 3.3% of hvCpGs to clustered protocadherin (*cPCDH*) genes on chromosome 5. This region comprises three clusters of protocadherin genes (*cPCDHα, cPCDHβ, cPCDHγ*), each containing many variable exons whose promoter choice is determined stochastically via differential methylation by DNA-methyltransferase 3 beta (DNMT3B) in early embryonic development(80, 81), resulting in the expression of distinct *cPCDH* isoforms of cell-surface proteins that are critical for establishing neuronal circuits(82). The *cPCDH* gene locus has also been found to be influenced by age(11, 83–85). Accordingly, although a minority (5%) of hvCpGs showed evidence of epigenetic drift in blood(11), these are enriched within the *cPCDH* locus relative to those that did not show evidence of epigenetic drift (Fisher's Exact Test (FET) p-value = 9.4×10^{-9}, OR = 4.02). Hypervariable methylation states at the *cPCDH* gene locus may therefore be driven by early developmental and/or aging effects. Noting that evCpGs and MZ twinning CpGs (Table 2) have also been reported to colocalise with this locus(53, 86), hvCpGs annotated to *cPCDH* genes were ~8.5-fold enriched for MZ twinning CpGs (FET p-value = 1.04×10^{-22}) and ~3-fold enriched for evCpGs (FET p-value = 1.6×10^{-3}) relative to hvCpGs that were not.

Association of hvCpGs with reported EWAS trait associations

To probe the potential functional role of hvCpGs, we analysed their overlap with traits reported in the epigenome-wide association studies (EWAS) catalogue (http://ewascatalog.org/). 86% of hvCpGs show significant associations (reported p-value < 1 x 10^{-4}) with one or more of 231 unique traits covered in the catalogue (Supplementary Table 9). However, compared to blood distribution-matched controls, a suitable comparator given that the majority of EWAS have been carried out in blood, we found that hvCpGs were enriched amongst CpGs associated with sex, Alzheimer's disease and inflammatory bowel disease only (Fig. 4B).

Noting that all sex-associated hvCpGs have top 5% methylation Beta variance in at least one of our 8 generated female-only and male-only datasets respectively ('Methods'), and that a similar proportion of SIV-CpGs are also associated with sex (23% of hvCpGs and 20% of the 3,089 SIV-CpGs considered in our study), we speculate that the association with sex may be a feature of variable methylation states established in early development. Amongst the 64 hvCpGs associated with Alzheimer's disease, 25 overlap previously identified SIV and/or ESS loci, 9 of which annotated to *CYP2E1*, a gene that has also been associated with Parkinson's disease and rheumatoid arthritis(32, 87). Amongst the 200 hvCpGs associated with inflammatory bowel disease, 87 overlap a SIV/ESS locus, 9 of which are annotated to *C21orf56*, a gene at which offspring methylation has been associated with maternal folate levels in pregnancy (88).

hvCpGs were notably depleted amongst age-related traits relative to distribution-matched controls (Fig. 4B), in agreement with our earlier findings that hvCpGs are largely stable with age (Supplementary Fig. 6). hvCpGs are also depleted amongst CpGs that are differentially methylated between buccal cells and peripheral blood mononucleocytes ('Tissue' in Fig. 4B), supporting the notion that hvCpGs may be established before cell differentiation and that the method used to identify the hvCpGs is robust to tissue-specific methylation variation.

Discussion

We have identified and characterised tissue- and ethnicity- independent hypervariable methylation states at CpGs covered by the 450k array. Our methodological approach was designed to be robust to dataset-specific drivers of methylation variability, including sex, age, cell type heterogeneity and technical artefacts. We identified 4,143 hvCpGs and found strong evidence that methylation states at many hvCpGs are likely to be established in the early embryo and are stable postnatally. Our analysis positions hvCpGs as tissue- and ethnicity- independent age-stable biomarkers of early stochastic and/or environmental effects on DNA methylation.

hvCpGs cover ~1% of the 450K array and were in the top 5% variable methylation states in an average of 13 distinct tissues and 7 ethnicities. Our study is not the first to investigate DNAm patterns in multiple tissues. Previous studies have identified CpGs that are differentially methylated between tissues (89–91); determined the extent to which variable methylation states in accessible tissues (such as blood) reflect those in inaccessible tissues such as brain (65, 90–93); compared methylation patterns between peripheral tissues (66, 94, 95); directly identified SIV loci using tissues derived from different germ layers (24–26, 28, 29); functionally characterised tissue-specific variably methylated regions(96); and examined the extent to which common drivers of methylation variation, such as genetics, age, sex and environment, are tissue-specific (8, 12, 15–18, 97, 98). The majority of these studies used a comparatively small number of tissues or cell-types, and few have used multi-tissue datasets from different ethnicities (15). To our knowledge, ours is the first study to explore the extent to which variably methylated CpGs are shared across diverse tissues and ethnicities in the human genome.

The majority of hvCpGs were associated with at least one mQTL suggesting that genetic effects influence methylation at these loci. Although data on the total proportion of methylation variance explained by all mQTLs associated with each hvCpG were not available, our comparison with mQTL-matched controls together with evidence of enrichment for sensitivity to periconceptional environment suggests that stochastic and/or environmental effects have a relatively large influence on methylation variability at hvCpGs. This is supported by evidence of methylation discordance between MZ twins, although we note that MZ discordance can be driven by *de novo* genetic mutations after MZ twinning events (99). A large proportion of hvCpGs show evidence of systemic interindividual variation (SIV), that is, intra-individual correlation in methylation across tissues derived from different germ layers. Whilst loci that covary across different tissue types are enriched for mQTL effects (16, 65, 66, 94), it has been suggested that SIV loci are putative human metastable epialleles with variable methylation states established before gastrulation that are influenced but not determined by genetic variation(28).

Our fetal multi-tissue analysis supports the notion that SIV at hvCpGs arises during early development and is likely not, for example, driven by post-natal environmental influences that act across many tissues. hvCpGs were also highly enriched for epigenetic supersimilarity loci and MZ twinning-associated CpGs, both of which have been linked to establishment of methylation in the cleavage stage pre-implantation embryo (27, 56). The degree of overlap between variably methylated regions in different cell types has also been linked to their common developmental origin(96). If this pattern holds true, it follows that stochastic and/or environmentally influenced variably methylated loci that are shared across a large number of diverse tissues are likely to have originated before germ-layer differentiation. Definitive proof of this would require an analysis of methylation variation at multiple stages in a sufficient number of pre-gastrulation embryos.

Although examination of EWAS trait associations revealed no evidence that hvCpGs are enriched for post-natal environmental effects, it is possible that cross-tissue and cross-ethnicity variable methylation states at some hvCpGs are influenced by later gestational or post-natal environmental effects. Such effects may act in addition to or independently of early environmental effects across multiple tissues, as has been suggested at the *VTRNA2-1* locus in the context of folate supplementation in pregnancy(100), maternal age at delivery(73), and smoking(101) .

An interesting feature of hvCpGs is their enrichment for intermediate methylation values relative to the array background. A previous read-level analysis of SIV loci from human embryos using reduced representation bisulphite sequencing data indicated that intermediate methylation states at SIV loci are driven by stochastic, cell variegation effects rather than allele-specific methylation (28), and this may also be the case for hvCpGs.

The association of hvCpGs with parent-of-origin-specific methylation and proximal ERV1 and ERVK elements is notable because these features have been linked to SIV-CpGs (28). This suggests that genomic regions targeted by epigenetic silencing or maintenance mechanisms during early embryonic reprogramming may be enriched for stochastic and/or environmentally influenced methylation variation. For example, it has been suggested that regions of PofOm may be vulnerable to stochastic or environmentally-sensitive loss of methylation on the usually-methylated allele or gain of methylation on the usually-unmethylated allele at a later time-point, leading to interindividual methylation variation (57, 71, 102). Similarly, certain IAP elements (a class of ERVK LTR retrotransposon) show methylation variation between isogenic mice (67, 68) that in several cases can be influenced by pre-natal environment (103–105). Whilst transposable elements are usually silenced to prevent insertion events from damaging the genome, recent evidence suggests that methylation variability at IAP elements is partly driven by low-affinity binding of *trans-a*cting Krüppel-associated box (KRAB)-containing zinc finger proteins (KZFPs) (69) and by sequence variation in KZFP-binding sites(69, 106). Whilst KZFPs are known to target TEs in humans (107, 108), the extent of their role in driving methylation variation is an ongoing area of research.

The large overlap between hvCpGs and 'high confidence' SIV-CpGs identified in at least two independent screens suggests that the identification of hvCpGs might constitute a high-powered method for detecting novel SIV loci. Supporting this, the largest SIV screen to date with 10 individuals was reported to be underpowered to detect the well-established SIV locus at the non-coding RNA gene *VTRNA2-1* (29) (represented by 10 hvCpGs), and we found that a 4-individual multi-tissue dataset analysed by van Baak *et al*.(27) had limited power to detect SIV loci. Another consideration is that SIV screens to date have used different sets of tissues. Since loci that covary between one pair of tissues do not necessarily covary between another pair (65), the enrichment for high confidence SIV loci (i.e. those reported in multiple independent SIV screens) might reflect the fact that methylation states at hvCpGs covary across a large number of tissues. Importantly, our analysis of a fetal multi-tissue dataset offers a strong validation of previously unreported SIV at hvCpGs.

Our analysis of EWAS trait associations revealed a moderate enrichment for hvCpGs amongst CpGs associated with Alzheimer's disease and inflammatory bowel disease. SIV loci have been linked to this and other disease outcomes including autism, cancer and obesity (27, 31, 109). For example, 10 hvCpGs overlap the *PAX8* gene which is a known SIV locus. *PAX8* methylation measured in peripheral blood of Gambian 2-year olds was recently shown to be correlated with thyroid volume and hormone levels in the same children in mid-childhood, and the latter was associated with changes in body fat and bone mineral density(110). This suggests that hvCpGs are interesting candidates for exploring how stochastic and/or environmentally influenced DNAm states established in early development might influence life-long health.

We identified hvCpGs that are variable in diverse ethnicities, raising the possibility that regions of hypervariable methylation may be a conserved feature in the human genome. It is also possible that there are ethnicity-specific regions of hypervariable methylation that would not have been captured in our analysis. Conserved variable methylation patterns established in the early embryo that are sensitive to early environment and that are able to influence gene expression might mediate a *predictive-adaptive-response* mechanism that senses the pre-natal environment in order to prime the developing embryo to its post-natal environment (75, 76). One hypothesis suggests that stochastic methylation states that are genetically hardwired into the human genome could provide a means of rapid adaptation to changing local environments on a scale much faster than is attainable through Darwinian evolution (112). Alternatively, stochastic methylation arising in early development independently of environmental factors may increase population fitness by expanding the range of phenotypes in a given generation (111) . Associations between genotype and methylation variance have been previously reported, for example at the putative metastable epiallele *PAX8*(36) at the master regulator of genomic imprinting *ZFP57* (27) and at several probes in the major histocompatibility complex (MHC) region associated with rheumatoid arthritis (113). Interestingly, 4% of hvCpGs are located within the MHC, representing an enrichment relative to the array background (FET p-value = 2.7×10^{-10}, OR = 1.7). Further analysis of genotype-methylation variance effects is required to

determine if this region, which contains a large amount of sequence variation and is implicated in many immune-mediated diseases (114), or others contain additional examples of genetically-driven phenotypic plasticity that is mediated by DNA methylation.

By selecting CpGs within the top 5% of methylation variance in at least 65% of datasets we were able to identify CpGs that were highly variable across multiple tissues and ethnicities. A comparison of CpGs identified using slightly different thresholds suggested that the set of hvCpGs is relatively insensitive to these parameters, but we note that the final set of hvCpGs is nevertheless dependent on the choice of thresholds.

Our method of adjusting for the first 10 PCs of variation increased power to detect consistent variable methylation states across datasets by reducing technical artefacts in each dataset, although some true biological variation may have been removed by doing so. It remains the case that we may not have controlled for all non-biological sources of variation within each dataset. As such, any remaining inter dataset differences due to unaccounted for technical variation and/or different pre-processing and normalisation steps already applied to public methylation data would result in a loss of power to detect hvCpGs. Conversely, if technical/normalisation issues were to cause a CpG to be in the top 5% of variance in one dataset, this CpG would be unlikely to be in the top 5% of variance across a majority of datasets. Inherent control of false positives arising from residual technical differences between datasets is therefore a strength of our approach.

We analysed methylation data covering 19 different tissues and 8 ethnicities. While these data were sufficiently powerful to identify several thousand hvCpGs, future analyses are likely to identify additional loci through the inclusion of larger datasets from diverse tissues and ethnicities as they become available. Furthermore, the vast majority of publicly available methylation datasets use the Illumina 450K array. Therefore, a major limitation of this study is that we were only able to analyse the small proportion of the methylome covered by this array, which has been found to miss a disproportionate amount of variable CpGs (29). However, we note that our method for identifying hypervariable CpGs can easily be applied to whole methylome sequencing data which is becoming increasingly available.

Through the joint analysis of methylation data from multiple tissues, we have identified a large set of hypervariable loci on the 450K array that are present across multiple tissues and ethnicities. Comparisons with a diverse range of data sources reveal that stochastic and/or environmentally responsive methylation states at these loci are likely to have been established in the early embryo and appear to be stable with age, making them interesting candidates for studying the developmental origins of life-long health and disease.

Data availability

The large majority of datasets analysed in this paper are in the public domain. GEO accession numbers and/or further details are provided in Supplementary Tables. A small number of analysed datasets have restricted access. Requests to access these should be submitted to the corresponding authors in the first instance with researcher access requiring an application to the relevant institutional review boards.

Acknowledgements

GoDMC meta-GWAS summary statistics were kindly provided by Jordana Bell. This data has now been published (Min et al, 2021).

Funding

This work was supported by the Biotechnology and Biological Sciences Research council (BB/M009513/1, BB/R01356X/1); the UK Medical Research Council (MR/M01424X/1, MC_PC_MR/RO20183/1 and MR/N006208/1); the Wellcome Trust/Royal Society (098386/Z/12/Z); the French National Research Agency (ANR OCEOADAPTO and ANR PAPUAEVOL 20-CE12-0003-01) and the Department of Biotechnology, Ministry of Science and Technology, India (BT/IN/DBT-MRC/DFID/24/GRC/2015–16). The fetal tissues supplied are part of the UCL-Imperial Baby Bio Bank collection. https://directory.biobankinguk.org

Conflict of Interest Disclosure

None declared.

References

1. Smith,Z.D. and Meissner,A. (2013) DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, **14**, 204–220.
2. Rakyan,V.K., Down,T.A., Balding,D.J. and Beck,S. (2011) Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, **12**, 529–541.
3. Lappalainen,T. and Greally,J.M. (2017) Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics*, **18**, 441–451.
4. Gaunt,T.R., Shihab,H.A., Hemani,G., Min,J.L., Woodward,G., Lyttleton,O., Zheng,J., Duggirala,A., McArdle,W.L., Ho,K., *et al.* (2016) Systematic identification of genetic influences on methylation across the human life course. *Genome Biology*, **17**, 61.
5. Bell,J.T., Pai,A.A., Pickrell,J.K., Gaffney,D.J., Pique-Regi,R., Degner,J.F., Gilad,Y. and Pritchard,J.K. (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology*, **12**, R10.
6. Jaffe,A.E. and Irizarry,R.A. (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biology*, **15**, R31.
7. Houseman,E.A., Kile,M.L., Christiani,D.C., Ince,T.A., Kelsey,K.T. and Marsit,C.J. (2016) Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, **17**, 259.
8. Singmann,P., Shem-Tov,D., Wahl,S., Grallert,H., Fiorito,G., Shin,S.-Y., Schramm,K., Wolf,P., Kunze,S., Baran,Y., *et al.* (2015) Characterization of whole-genome autosomal differences of DNA methylation between men and women. *Epigenetics & Chromatin 2015 8:1*, **8**, 1–13.
9. Yousefi,P., Huen,K., Davé,V., Barcellos,L., Eskenazi,B. and Holland,N. (2015) Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. *BMC Genomics*, **16**, 911.
10. Horvath,S. (2013) DNA methylation age of human tissues and cell types. *Genome Biol*, **14**, 1–20.
11. Slieker,R.C., van Iterson,M., Luijk,R., Beekman,M., Zhernakova,D. v, Moed,M.H., Mei,H., van Galen,M., Deelen,P. and Bonder,M.J. (2016) Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol*, **17**, 1–13.
12. Busche,S., Shao,X., Caron,M., Kwan,T., Allum,F., Cheung,W.A., Ge,B., Westfall,S., Simon,M.-M. and Barrett,A. (2015) Population whole-genome bisulfite sequencing across two tissues highlights the environment as the principal source of human methylome variation. *Genome Biol*, **16**, 1–18.
13. Hannon,E., Knox,O., Sugden,K., Burrage,J., Wong,C.C.Y., Belsky,D.W., Corcoran,D.L., Arseneault,L., Moffitt,T.E., Caspi,A., *et al.* (2018) Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLOS Genetics*, **14**, e1007544–e1007544.
14. Czamara,D., Eraslan,G., Page,C.M., Lahti,J., Lahti-Pulkkinen,M., Hämäläinen,E., Kajantie,E., Laivuori,H., Villa,P.M., Reynolds,R.M., *et al.* (2019) Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. *Nature Communications*, **10**.
15. Smith,A.K., Kilaru,V., Kocak,M., Almli,L.M., Mercer,K.B., Ressler,K.J., Tylavsky,F.A. and Conneely,K.N. (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. *BMC Genomics*, **15**, 145.
16. Lin,D., Chen,J., Perrone-Bizzozero,N., Bustillo,J.R., Du,Y., Calhoun,V.D. and Liu,J. (2018) Characterization of cross-tissue genetic-epigenetic effects and their patterns in schizophrenia. *Genome Medicine*, **10**, 13.
17. Day,K., Waite,L.L., Thalacker-Mercer,A., West,A., Bamman,M.M., Brooks,J.D., Myers,R.M. and Absher,D. (2013) Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. *Genome Biology*, **14**.
18. Slieker,R.C., Relton,C.L., Gaunt,T.R., Slagboom,P.E. and Heijmans,B.T. (2018) Age-related DNA methylation changes are tissue-specific with ELOVL2 promoter methylation as exception. *Epigenetics and Chromatin*, **11**.
19. Bibikova,M., Barnes,B., Tsan,C., Ho,V., Klotzle,B., Le,J.M., Delano,D., Zhang,L., Schroth,G.P., Gunderson,K.L., *et al.* (2011) High density DNA methylation array with single CpG site resolution. *Genomics*, **98**, 288–295.

20. Price,E.M. and Robinson,W.P. (2018) Adjusting for Batch Effects in DNA Methylation Microarray Data, a Lesson Learned. *Frontiers in Genetics*, **0**, 83.
21. Leek,J.T., Scharpf,R.B., Bravo,H.C., Simcha,D., Langmead,B., Johnson,W.E., Geman,D., Baggerly,K. and Irizarry,R.A. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, **11**, 733–739.
22. Harper,K.N., Peters,B.A. and Gamble,M. v. (2013) Batch effects and pathway analysis: Two potential perils in cancer studies involving DNA methylation array analysis. *Cancer Epidemiology Biomarkers and Prevention*, **22**, 1052–1060.
23. Sugden,K., Hannon,E.J., Arseneault,L., Belsky,D.W., Corcoran,D.L., Fisher,H.L., Houts,R.M., Kandaswamy,R., Moffitt,T.E., Poulton,R., *et al.* (2020) Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. *Patterns*, **1**.
24. Waterland,R.A., Kellermayer,R., Laritsky,E., Rayco-Solon,P., Harris,R.A., Travisano,M., Zhang,W., Torskaya,M.S., Zhang,J., Shen,L., *et al.* (2010) Season of Conception in Rural Gambia Affects DNA Methylation at Putative Human Metastable Epialleles. *PLoS Genetics*, **6**, e1001252.
25. Silver,M.J., Kessler,N.J., Hennig,B.J., Dominguez-Salas,P., Laritsky,E., Baker,M.S., Coarfa,C., Hernandez-Vargas,H., Castelino,J.M., Routledge,M.N., *et al.* (2015) Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biology*, **16**.
26. Alan Harris,R., Nagy-Szakal,D. and Kellermayer,R. (2013) Human metastable epiallele candidates link to common disorders. *Epigenetics*, **8**, 157–163.
27. van Baak,T.E., Coarfa,C., Dugué,P.-A., Fiorito,G., Laritsky,E., Baker,M.S., Kessler,N.J., Dong,J., Duryea,J.D., Silver,M.J., *et al.* (2018) Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biology*, **19**, 2.
28. Kessler,N.J., Waterland,R.A., Prentice,A.M. and Silver,M.J. (2018) Establishment of environmentally sensitive DNA methylation states in the very early human embryo. *Sci Adv*, **4**, eaat2624.
29. Gunasekara,C.J., Scott,C.A., Laritsky,E., Baker,M.S., MacKay,H., Duryea,J.D., Kessler,N.J., Hellenthal,G., Wood,A.C., Hodges,K.R., *et al.* (2019) A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biology*, **20**, 105.
30. Finer,S., Iqbal,M.S., Lowe,R., Ogunkolade,B.W., Pervin,S., Mathews,C., Smart,M., Alam,D.S. and Hitman,G.A. (2016) Is famine exposure during developmental life in rural Bangladesh associated with a metabolic and epigenetic signature in young adulthood? A historical cohort study. *BMJ Open*, **6**, e011768.
31. Kühnen,P., Handke,D., Waterland,R.A., Hennig,B.J., Silver,M., Fulford,A.J., Dominguez-Salas,P., Moore,S.E., Prentice,A.M., Spranger,J., *et al.* (2016) Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity. *Cell Metabolism*, **24**, 502–509.
32. Mok,A., Rhead,B., Holingue,C., Shao,X., Quach,H.L., Quach,D., Sinclair,E., Graf,J., Imboden,J., Link,T., *et al.* (2018) Hypomethylation of CYP2E1 and DUSP22 Promoters Associated With Disease Activity and Erosive Disease Among Rheumatoid Arthritis Patients. *Arthritis and Rheumatology*, **70**, 528–536.
33. Zhu,Y., Mordaunt,C.E., Yasui,D.H., Marathe,R., Coulson,R.L., Dunaway,K.W., Jianu,J.M., Walker,C.K., Ozonoff,S., Hertz-Picciotto,I., *et al.* (2019) Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Human Molecular Genetics*, **28**, 2659–2674.
34. Sanchez-Mut,J. v, Heyn,H., Silva,B.A., Dixsaut,L., Garcia-Esparcia,P., Vidal,E., Sayols,S., Glauser,L., Monteagudo-Sánchez,A., Perez-Tur,J., *et al.* (2018) PM20D1 is a quantitative trait locus associated with Alzheimer's disease. *Nature Medicine*, **24**, 598–603.
35. Young,J.I., Sivasankaran,S.K., Wang,L., Ali,A., Mehta,A., Davis,D.A., Dykxhoorn,D.M., Petito,C.K., Beecham,G.W., Martin,E.R., *et al.* (2019) Genome-wide brain DNA methylation analysis suggests epigenetic reprogramming in Parkinson disease. *Neurology: Genetics*, **5**.
36. Candler,T., Kessler,N., Gunasekara,C., Ward,K., James,P., Laritsky,E., Baker,M., Dyer,R., Elango,R. and Jeffries,D. (2021) DNA methylation at a nutritionally sensitive region of the PAX8 gene is associated with thyroid volume and function in Gambian children. *Science Advances*, **7**, eabj1561.
37. Gunasekara,C.J. and Waterland,R.A. (2019) A new era for epigenetic epidemiology. *Epigenomics*, **11**, 1647–1649.

38. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, **41**, D991–D995.
39. Silva,T.C., Colaprico,A., Olsen,C., D'Angelo,F., Bontempi,G., Ceccarelli,M. and Noushmehr,H. (2016) TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Res*, **5**, 1542.
40. Mounir,M., Lucchetta,M., Silva,T.C., Olsen,C., Bontempi,G., Chen,X., Noushmehr,H., Colaprico,A. and Papaleo,E. (2019) New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Comput Biol*, **15**, e1006701–e1006701.
41. Colaprico,A., Silva,T.C., Olsen,C., Garofano,L., Cava,C., Garolini,D., Sabedot,T.S., Malta,T.M., Pagnotta,S.M., Castiglioni,I., *et al.* (2016) TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research*, **44**, e71–e71.
42. Davis,S. and Meltzer,P.S. (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.
43. Moran,S., Arribas,C. and Esteller,M. (2016) Validation of a DNA methylation microarray for 850,000 CpG sites of the human genome enriched in enhancer sequences. *Epigenomics*, **8**, 389–399.
44. Chandak,G.R., Silver,M.J., Saffari,A., Lillycrop,K.A., Shrestha,S., Sahariah,S.A., di Gravio,C., Goldberg,G., Tomar,A.S., Betts,M., *et al.* (2017) Protocol for the EMPHASIS study; epigenetic mechanisms linking maternal pre-conceptional nutrition and children's health in India and Sub-Saharan Africa. *BMC Nutrition*, **3**, 81.
45. Brucato,N., Fernandes,V., Mazières,S., Kusuma,P., Cox,M.P., Ng'ang'a,J.W., Omar,M., Simeone-Senelle,M.-C., Frassati,C., Alshamali,F., *et al.* (2018) The Comoros Show the Earliest Austronesian Gene Flow into the Swahili Corridor. *Am J Hum Genet*, **102**, 58–68.
46. Morris,T.J., Butcher,L.M., Feber,A., Teschendorff,A.E., Chakravarthy,A.R., Wojdacz,T.K. and Beck,S. (2014) ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*, 10.1093/bioinformatics/btt684.
47. Nordlund,J., Bäcklin,C.L., Wahlberg,P., Busche,S., Berglund,E.C., Eloranta,M.-L., Flaegstad,T., Forestier,E., Frost,B.-M., Harila-Saari,A., *et al.* (2013) Genome-wide signatures of differential DNA methylation in pediatric acute lymphoblastic leukemia. *Genome Biology*, **14**, r105.
48. Zhou,W., Laird,P.W. and Shen,H. (2017) Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res*, **45**, e22–e22.
49. Teschendorff,A.E., Marabita,F., Lechner,M., Bartlett,T., Tegner,J., Gomez-Cabrero,D. and Beck,S. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, **29**, 189–196.
50. Bose,M., Wu,C., Pankow,J.S., Demerath,E.W., Bressler,J., Fornage,M., Grove,M.L., Mosley,T.H., Hicks,C., North,K., *et al.* (2014) Evaluation of microarray-based DNA methylation measurement using technical replicates: The atherosclerosis risk in communities (ARIC) study. *BMC Bioinformatics*, **15**, 312.
51. Min,J.L., Hemani,G., Hannon,E., Dekkers,K.F., Castillo-Fernandez,J., Luijk,R., Carnero-Montoro,E., Lawson,D.J., Burrows,K. and Suderman,M. (2021) Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*, **53**, 1311–1321.
52. Bonilla,C., Bertoni,B., Min,J.L., Hemani,G., Consortium,G. of D.N.A.M. and Elliott,H.R. (2021) Investigating DNA methylation as a potential mediator between pigmentation genes, pigmentary traits and skin cancer. *Pigment Cell Melanoma Res*, **34**, 892–904.
53. Planterose Jiménez,B., Liu,F., Caliebe,A., Montiel González,D., Bell,J.T., Kayser,M. and Vidaki,A. (2021) Equivalent DNA methylation variation between monozygotic co-twins and unrelated individuals reveals universal epigenetic inter-individual dissimilarity. *Genome Biology*, **22**, 18.
54. Flanagan,J.M., Brook,M.N., Orr,N., Tomczyk,K., Coulson,P., Fletcher,O., Jones,M.E., Schoemaker,M.J., Ashworth,A., Swerdlow,A., *et al.* (2015) Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. *Cancer Epidemiology Biomarkers and Prevention*, **24**, 221–229.
55. Grundberg,E., Meduri,E., Sandling,J.K., Hedman,A.K., Keildson,S., Buil,A., Busche,S., Yuan,W., Nisbet,J., Sekowska,M., *et al.* (2013) Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *Am J Hum Genet*, **93**, 876–90.

56. van Dongen,J., Gordon,S.D., McRae,A.F., Odintsova,V. v., Mbarek,H., Breeze,C.E., Sugden,K., Lundgren,S., Castillo-Fernandez,J.E., Hannon,E., *et al.* (2021) Identical twins carry a persistent epigenetic signature of early genome programming. *Nature Communications*, **12**, 5618.
57. Silver,M.J., Saffari,A., Kessler,N.J., Chandak,G.R., Fall,C.H.D., Issarapu,P., Dedaniya,A., Betts,M., Moore,S.E., Routledge,M.N., *et al.* (2022) Environmentally sensitive hotspots in the methylome of the early human embryo. *Elife*, **11**, e72031.
58. Zink,F., Magnusdottir,D.N., Magnusson,O.T., Walker,N.J., Morris,T.J., Sigurdsson,A., Halldorsson,G.H., Gudjonsson,S.A., Melsted,P., Ingimundardottir,H., *et al.* (2018) Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nature Genetics*, **50**, 1542–1552.
59. Lokk,K., Modhukur,V., Rajashekar,B., Märtens,K., Mägi,R., Kolde,R., Koltšina,M., Nilsson,T.K., Vilo,J., Salumets,A., *et al.* (2014) DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol*, 10.1186/gb-2014-15-4-r54.
60. Min,J.L., Hemani,G., Smith,G.D., Relton,C. and Suderman,M. (2018) Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics*, **34**, 3983.
61. Ernst,J. and Kellis,M. (2012) ChromHMM: Automating chromatin-state discovery and characterization. *Nature Methods*, **9**, 215–216.
62. Roadmap Epigenomics Consortium, Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J., *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, 10.1038/nature14248.
63. Phipson,B., Maksimovic,J. and Oshlack,A. (2016) MissMethyl: An R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*, **32**.
64. Fraga,M.F., Ballestar,E., Paz,M.F., Ropero,S., Setien,F., Ballestar,M.L., Heine-Suñer,D., Cigudosa,J.C., Urioste,M. and Benitez,J. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences*, **102**, 10604–10609.
65. Hannon,E., Lunnon,K., Schalkwyk,L. and Mill,J. (2015) Interindividual methylomic variation across blood, cortex, and cerebellum: Implications for epigenetic studies of neurological and neuropsychiatric phenotypes. *Epigenetics*, **10**, 1024–1032.
66. Islam,S.A., Goodman,S.J., MacIsaac,J.L., Obradović,J., Barr,R.G., Boyce,W.T. and Kobor,M.S. (2019) Integration of DNA methylation patterns and genetic variation in human pediatric tissues help inform EWAS design and interpretation. *Epigenetics Chromatin*, **12**, 1–18.
67. Rakyan,V.K., Blewitt,M.E., Druker,R., Preis,J.I. and Whitelaw,E. (2002) Metastable epialleles in mammals. *Trends in Genetics*, **18**, 348–351.
68. Kazachenka,A., Bertozzi,T.M., Sjoberg-Herrera,M.K., Walker,N., Gardner,J., Gunning,R., Pahita,E., Adams,S., Adams,D. and Ferguson-Smith,A.C. (2018) Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell*, **175**, 1259-1271.e13.
69. Bertozzi,T.M., Elmer,J.L., Macfarlan,T.S. and Ferguson-Smith,A.C. (2020) KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. *Proceedings of the National Academy of Sciences*, 10.1073/pnas.2017053117.
70. Elmer,J.L., Hay,A.D., Kessler,N.J., Bertozzi,T.M., Ainscough,E.A.C. and Ferguson-Smith,A.C. (2021) Genomic properties of variably methylated retrotransposons in mouse. *Mob DNA*, **12**.
71. Monk,D., Mackay,D.J.G., Eggermann,T., Maher,E.R. and Riccio,A. (2019) Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews Genetics*, **20**, 235–248.
72. Carpenter,B.L., Remba,T.K., Thomas,S.L., Madaj,Z., Brink,L., Tiedemann,R.L., Odendaal,H.J. and Jones,P.A. (2021) Oocyte age and preconceptual alcohol use are highly correlated with epigenetic imprinting of a noncoding RNA (nc886). *Proceedings of the National Academy of Sciences*, **118**.
73. Markunas,C.A., Wilcox,A.J., Xu,Z., Joubert,B.R., Harlid,S., Panduri,V., Håberg,S.E., Nystad,W., London,S.J., Sandler,D.P., *et al.* (2016) Maternal age at delivery is associated with an epigenetic signature in both newborns and adults. *PLoS ONE*, **11**.
74. Gluckman,P.D., Hanson,M.A. and Low,F.M. (2011) The role of developmental plasticity and epigenetics in human health. *Birth Defects Research Part C - Embryo Today: Reviews*, **93**, 12–18.
75. Low,F.M., Gluckman,P.D. and Hanson,M.A. (2012) Developmental Plasticity, Epigenetics and Human Health. *Evolutionary Biology*, 10.1007/s11692-011-9157-0.

76. Fleming,T.P., Watkins,A.J., Velazquez,M.A., Mathers,J.C., Prentice,A.M., Stephenson,J., Barker,M., Saffery,R., Yajnik,C.S., Eckert,J.J., *et al.* (2018) Origins of lifetime health around the time of conception: causes and consequences. *The Lancet*, **391**, 1842–1852.
77. Moore,S.E., Cole,T.J., Collinson,A.C., Poskitt,E.M.E., McGregor,I.A. and Prentice,A.M. (1999) Prenatal or early postnatal events predict infectious deaths in young adulthood in rural Africa. *International Journal of Epidemiology*, 10.1093/ije/28.6.1088.
78. Dominguez-Salas,P., Moore,S.E., Cole,D., da Costa,K.A., Cox,S.E., Dyer,R.A., Fulford,A.J.C., Innis,S.M., Waterland,R.A., Zeisel,S.H., *et al.* (2013) DNA methylation potential: Dietary intake and blood concentrations of one-carbon metabolites and cofactors in rural African women. *American Journal of Clinical Nutrition*, **97**, 1217–1227.
79. James,P.T., Dominguez-Salas,P., Hennig,B.J., Moore,S.E., Prentice,A.M. and Silver,M.J. (2019) Maternal One-Carbon Metabolism and Infant DNA Methylation between Contrasting Seasonal Environments: A Case Study from The Gambia. *Current Developments in Nutrition*, **3**.
80. el Hajj,N., Dittrich,M. and Haaf,T. (2017) Epigenetic dysregulation of protocadherins in human disease. *Seminars in Cell and Developmental Biology*, **69**, 172–182.
81. Toyoda,S., Kawaguchi,M., Kobayashi,T., Tarusawa,E., Toyama,T., Okano,M., Oda,M., Nakauchi,H., Yoshimura,Y., Sanbo,M., *et al.* (2014) Developmental epigenetic modification regulates stochastic expression of clustered Protocadherin genes, generating single neuron diversity. *Neuron*, **82**, 94–108.
82. Flaherty,E. and Maniatis,T. (2020) The role of clustered protocadherins in neurodevelopment and neuropsychiatric diseases. *Current Opinion in Genetics and Development*, **65**, 144–150.
83. Salpea,P., Russanova,V.R., Hirai,T.H., Sourlingas,T.G., Sekeri-Pataryas,K.E., Romero,R., Epstein,J. and Howard,B.H. (2012) Postnatal development- and age-related changes in DNA-methylation patterns in the human genome. *Nucleic Acids Research*, **40**, 6477–6494.
84. McClay,J.L., Aberg,K.A., Clark,S.L., Nerella,S., Kumar,G., Xie,L.Y., Hudson,A.D., Harada,A., Hultman,C.M., Magnusson,P.K.E., *et al.* (2014) A methylome-wide study of aging using massively parallel sequencing of the methyl-CpG-enriched genomic fraction from blood in over 700 subjects. *Human Molecular Genetics*, **23**, 1175–1185.
85. Kim,S., Wyckoff,J., Morris,A.-T., Succop,A., Avery,A., Duncan,G.E. and Jazwinski,S.M. (2018) DNA methylation associated with healthy aging of elderly twins. *GeroScience 2018 40:5*, **40**, 469–484.
86. van Dongen,J., Nivard,M.G., Willemsen,G., Hottenga,J.J., Helmer,Q., Dolan,C. v, Ehli,E.A., Davies,G.E., van Iterson,M., Breeze,C.E., *et al.* (2016) Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nature Communications*, **7**.
87. Kaut,O., Schmitt,I. and Wüllner,U. (2012) Genome-scale methylation analysis of Parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1. *Neurogenetics*, **13**, 87–91.
88. Amarasekera,M., Martino,D., Ashley,S., Harb,H., Kesper,D., Strickland,D., Saffery,R. and Prescott,S.L. (2014) Genome-wide DNA methylation profiling identifies a folate-sensitive region of differential methylation upstream of ZFP57-imprinting regulator in humans. *The FASEB Journal*, **28**, 4068–4076.
89. Byun,H.M., Siegmund,K.D., Pan,F., Weisenberger,D.J., Kanel,G., Laird,P.W. and Yang,A.S. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Human Molecular Genetics*, **18**, 4808–4817.
90. Slieker,R.C., Bos,S.D., Goeman,J.J., Bovée,J.V., Talens,R.P., van der Breggen,R., Suchiman,H.E.D., Lameijer,E.W., Putter,H., van den Akker,E.B., *et al.* (2013) Identification and systematic annotation of tissue-specific differentially methylated regions using the Illumina 450k array. *Epigenetics and Chromatin*, **6**.
91. Davies,M.N., Volta,M., Pidsley,R., Lunnon,K., Dixit,A., Lovestone,S., Coarfa,C., Harris,R.A., Milosavljevic,A., Troakes,C., *et al.* (2012) Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol*, **13**, R43–R43.
92. Walton,E., Hass,J., Liu,J., Roffman,J.L., Bernardoni,F., Roessner,V., Kirsch,M., Schackert,G., Calhoun,V. and Ehrlich,S. (2016) Correspondence of DNA methylation between blood and brain tissue and its application to schizophrenia research. *Schizophrenia Bulletin*, **42**, 406–414.
93. Braun,P.R., Han,S., Hing,B., Nagahama,Y., Gaul,L.N., Heinzman,J.T., Grossbach,A.J., Close,L., Dlouhy,B.J., Howard,M.A., *et al.* (2019) Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals. *Translational Psychiatry*, **9**, 1–10.

94. Hannon,E., Mansell,G., Walker,E., Nabais,M.F., Burrage,J., Kepa,A., Best-Lane,J., Rose,A., Heck,S., Moffitt,T.E., *et al.* (2021) Assessing the co-variability of DNA methylation across peripheral cells and tissues: Implications for the interpretation of findings in epigenetic epidemiology. *PLOS Genetics*, **17**, e1009443.
95. Jiang,R., Jones,M.J., Chen,E., Neumann,S.M., Fraser,H.B., Miller,G.E. and Kobor,M.S. (2015) Discordance of DNA methylation variance between two accessible human tissues. *Scientific Reports*, **5**, 8257.
96. Garg,P., Joshi,R.S., Watson,C. and Sharp,A.J. (2018) A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. *PLOS Genetics*, **14**, e1007707–e1007707.
97. Gordon,L., Joo,J.E., Powell,J.E., Ollikainen,M., Novakovic,B., Li,X., Andronikos,R., Cruickshank,M.N., Conneely,K.N., Smith,A.K., *et al.* (2012) Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Research*, **22**, 1395–1406.
98. Do,C., Lang,C.F., Lin,J., Darbary,H., Krupska,I., Gaba,A., Petukhova,L., Vonsattel,J.P., Gallagher,M.P., Goland,R.S., *et al.* (2016) Mechanisms and Disease Associations of Haplotype-Dependent Allele-Specific DNA Methylation. *American Journal of Human Genetics*, **98**, 934–955.
99. Vadgama,N., Pittman,A., Simpson,M., Nirmalananthan,N., Murray,R., Yoshikawa,T., de Rijk,P., Rees,E., Kirov,G. and Hughes,D. (2019) De novo single-nucleotide and copy number variation in discordant monozygotic twins reveals disease-related genes. *European Journal of Human Genetics*, **27**, 1121–1133.
100. Richmond,R.C., Sharp,G.C., Herbert,G., Atkinson,C., Taylor,C., Bhattacharya,S., Campbell,D., Hall,M., Kazmi,N., Gaunt,T., *et al.* (2018) The long-term impact of folic acid in pregnancy on offspring DNA methylation: Follow-up of the Aberdeen folic acid supplementation trial (AFAST). *International Journal of Epidemiology*, **47**, 928–937.
101. Ambatipudi,S., Cuenin,C., Hernandez-Vargas,H., Ghantous,A., le Calvez-Kelm,F., Kaaks,R., Barrdahl,M., Boeing,H., Aleksandrova,K. and Trichopoulou,A. (2016) Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. *Epigenomics*, **8**, 599–618.
102. James,P., Sajjadi,S., Tomar,A.S., Saffari,A., Fall,C.H.D., Prentice,A.M., Shrestha,S., Issarapu,P., Yadav,D.K., Kaur,L., *et al.* (2018) Candidate genes linking maternal nutrient exposure to offspring health via DNA methylation: A review of existing evidence in humans with specific focus on one-carbon metabolism. *International Journal of Epidemiology*, 10.1093/ije/dyy153.
103. Wolff,G.L., Kodell,R.L., Moore,S.R. and Cooney,C.A. (1998) Maternal epigenetics and methyl supplements affect agouti gene expression in A vy /a mice.
104. Cooney,C.A., Dave,A.A. and Wolff,G.L. (2002) Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. In *Journal of Nutrition*.
105. Dolinoy,D.C., Huang,D. and Jirtle,R.L. (2007) Maternal nutrient supplementation counteracts bisphenol A-induced DNA hypomethylation in early development. *Proceedings of the National Academy of Sciences*, **104**, 13056–13061.
106. Costello,K.R., Leung,A., Trac,C., Lee,M., Basam,M., Pospisilik,J.A. and Schones,D.E. (2021) Mechanisms regulating interindividual epigenetic variability at transposable elements. *bioRxiv*.
107. Imbeault,M., Helleboid,P.Y. and Trono,D. (2017) KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature*, **543**, 550–554.
108. Shen,P., Xu,A., Hou,Y., Wang,H., Gao,C., He,F. and Yang,D. (2021) Conserved paradoxical relationships among the evolutionary, structural and expressional features of KRAB zinc-finger proteins reveal their special functional characteristics. *BMC Molecular and Cell Biology 2021 22:1*, **22**, 1–15.
109. Zhu,Y., Mordaunt,C.E., Yasui,D.H., Marathe,R., Coulson,R.L., Dunaway,K.W., Jianu,J.M., Walker,C.K., Ozonoff,S., Hertz-Picciotto,I., *et al.* (2019) Placental DNA methylation levels at CYP2E1 and IRS2 are associated with child outcome in a prospective autism study. *Human Molecular Genetics*, **28**, 2659–2674.
110. Candler,T., Kühnen,P., Prentice,A.M. and Silver,M. (2019) Epigenetic regulation of POMC; implications for nutritional programming, obesity and metabolic disease. *Frontiers in Neuroendocrinology*, **54**.
111. Ballouz,S., Pena,M.T., Knight,F.M., Adams,L.B. and Gillis,J.A. (2019) The transcriptional legacy of developmental stochasticity. *bioRxiv*, 10.1101/2019.12.11.873265.

112. Feinberg,A.P. and Irizarry,R.A. (2010) Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proceedings of the National Academy of Sciences*, **107**, 1757–1764.
113. Liu,Y., Aryee,M.J., Padyukov,L., Fallin,M.D., Hesselberg,E., Runarsson,A., Reinius,L., Acevedo,N., Taub,M., Ronninger,M., *et al.* (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol*, **31**, 142–7.
114. Matzaraki,V., Kumar,V., Wijmenga,C. and Zhernakova,A. (2017) The MHC locus and genetic susceptibility to autoimmune and infectious diseases. *Genome Biology*, **18**, 76.

Figure legends

Figure 1. Identification of tissue- and ethnicity- independent hypervariable CpGs ('hvCpGs'). A) CpGs with methylation Beta variance within the top 20%, 19%, 18%... 1% of variable CpGs were first extracted from each of the 30 methylation datasets used in this study. The intersection of these CpGs was then taken over an increasing proportion of datasets, requiring each CpG to be present in a minimum of 15 of the datasets analysed. The heatmap shows the number of CpGs within the top *i* % of variable sites by methylation Beta variance ('variance threshold') overlapping at least *j* % of datasets. To identify hypervariable CpGs ('hvCpGs'), we set a threshold at *i,j* = [5,65], marked by the orange box. **B)** Bar charts showing the proportion of the set of 4,330 hvCpGs identified using *i,j* = [5,65] that have top 5% methylation Beta variance in ≥ n ethnicities (left) and tissues (right). See Supplementary Table 4 for groupings of datasets by tissue type and ethnicity.

Figure 2. Genetic effects at hvCpGs using mQTL data from a large meta GWAS in blood (Min et al. 2021). A) The relationship between hypervariability and the proportion of CpGs with at least one mQTL association (top) and the mean mQTL effect size (bottom). Coloured curves represent CpGs with top 5% (orange), 10% (red) and 20% (blue) methylation Beta variance in at least *x%* of datasets. **B**) mQTL effects at hvCpGs**.** Left: the proportion of hvCpGs that are associated with *n* mQTLs. Right: the distribution of the mean % variance explained by mQTLs at 3,722 hvCpGs that are associated with at least one mQTL. **C)** Median methylation Beta variance at 3,722 hvCpGs overlapping the 'Blood_Cauc' dataset (orange) and corresponding controls matched on number of mQTL associations and mean % explained by mQTL ('mQTL-matched controls', Table1; Supplementary Fig. 4), in each dataset. See Supplementary Tables 1-3 for further details on tissues and ethnicities. Error bars in A and C are bootstrapped 95% confidence intervals. Note, error bars in C are very small.

Figure 3. hvCpGs are enriched for loci and genomic features linked to variable methylation establishment in early development. A) The proportion of 3,566 hvCpGs (y-axis) vs corresponding distribution-matched controls (x-axis) covered in the 'Blood_Cauc' dataset that overlap 3,089 SIV-CpGs, 1,217 ESS CpGs identified by van Baak *et al.* (2018), 728 'MZ twinning' CpGs identified by van Dongen *et al.* (2021) and 732 PofOm CpGs identified by Zink *et al.* (2018). **B)** The proportion of SIV-CpGs, ESS CpGs, MZ twinning CpGs and PofOm CpGs that are hvCpGs. SIV-CpGs identified in at least two or three independent screens were also included in this plot. **C)** Inter-germ layer correlations at hvCpGs using a fetal multi-tissue dataset that comprises methylation data from 10 individuals with endoderm- and ectoderm- derived tissues, 9 individuals with endoderm- and mesoderm- derived tissues and 8 individuals with mesoderm- and ectoderm- derived tissues (see Supplementary Table 7). **Left**: The distribution of average inter-germ layer correlations at 3,878 hvCpGs (orange) and 372,571 array background CpGs (excluding previously identified SIV CpGs and hvCpGs) (dark grey) covered in the fetal multi-tissue dataset. **Top Right**: Interindividual variation at 3,878 hvCpGs (orange), 4,076 previously identified SIV loci (blue) covered in the fetal multi-tissue dataset, and 372,571 array background CpGs (see 'Methods' for definition of interindividual variation). **Bottom Right:** Comparison of average inter-germ layer correlations at hvCpGs, SIV- CpGs and array background CpGs, stratified by interindividual variation. Each point indicates the median average inter-germ layer correlation for those CpGs with interindividual variation falling within each bound specified on the x-axis. **D)** The proportion of 3,566 hvCpGs, distribution-matched controls and array background CpGs that are ≤ *x* bp from the nearest ERV1 and ERVK transposable elements determined by RepeatMasker. Error bars in all panels are bootstrapped 95% confidence intervals. SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, PofOm = parent-of-origin-specific methylation.

Figure 4. Functional annotation of hvCpGs. **A)** Gene ontology term enrichment analysis at hvCpGs. Vertical line indicates a significance threshold of FDR < 0.05. **B)** EWAS trait enrichment of hvCpGs relative to blood-distribution controls (Table 1) for traits overlapping at least 1% of hvCpGs (see 'Methods' and Supplementary Table 9 for further details). X-axis gives enrichment odds ratios and bar colour gives Fisher's Exact Test (FET) significance p-values. Shown are the 10 traits with FET p-value ≤ 0.01.

Tables -remember to change numbers

Table 1. Main CpG sets used in this study.

CpG set	n	Notes
hvCpGs	4143	CpGs within top 5% methylation Beta variance in at least 65% datasets in which the CpG is covered, requiring the hvCpG to be covered in at least 15 datasets and to be reported as technically reliable.
array background	406306	CpGs covered in at least 15 of the 30 datasets used in this study.
distribution-matched controls	3566	Array background CpGs with similar methylation Beta distributions to hvCpGs in the 'Blood_Cauc' dataset, requiring each control CpG to be technically reliable.
de-clustered hvCpGs	2640	A set of hvCpGs in which no CpGs is within 4kb of another CpG.
mQTL-matched controls	3722	CpGs reported by the GoDMC meta-GWAS (51) with the same number of mQTL associations and similar mean % variance explained by an mQTL, requiring each control CpG to be present in at least as many datasets as the hvCpG.

Table 2. Published CpG sets used in this study

CpG set	Description	n. CpGs overlapping array background	Reference
SIV	Interindividual methylation variation with concordant methylation across tissues derived from different germ layers within a given individual. See Supplementary Table 8.	3089	Harris *et al.,* van Baak *et al*., Kessler *et al.,* Gunasekara *et al.* (26-29)
ESS	Greater-than-expected methylation similarity between MZ co-twins	1217	van Baak *et al.* (27)
MZ twinning CpGs	Probes differentially methylated between MZ and DZ twins.	728	van Dongen *et al.* (56)
evCpGs	MZ co-twin methylation discordance that is equivalent to methylation discordance between unrelated individuals in whole blood. A subset of these replicated in adipose tissue.	317 (blood) 145 (blood & adipose)	Planterose Jimé nez *et al.* (53)
SoC	CpGs at which methylation is associated with season of conception in Gambian children.	242	Silver *et al.* (57)
PofOm	Regions of parent-of-origin-specific methylation identified in peripheral blood from Icelandic individuals.	732 CpGs in 116 PofOm regions	Zink *et al.* (58)

SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, evCpGs = equivalently variable CpGs, SoC = season-of-conception, PofOm = parent-of-origin-specific methylation.

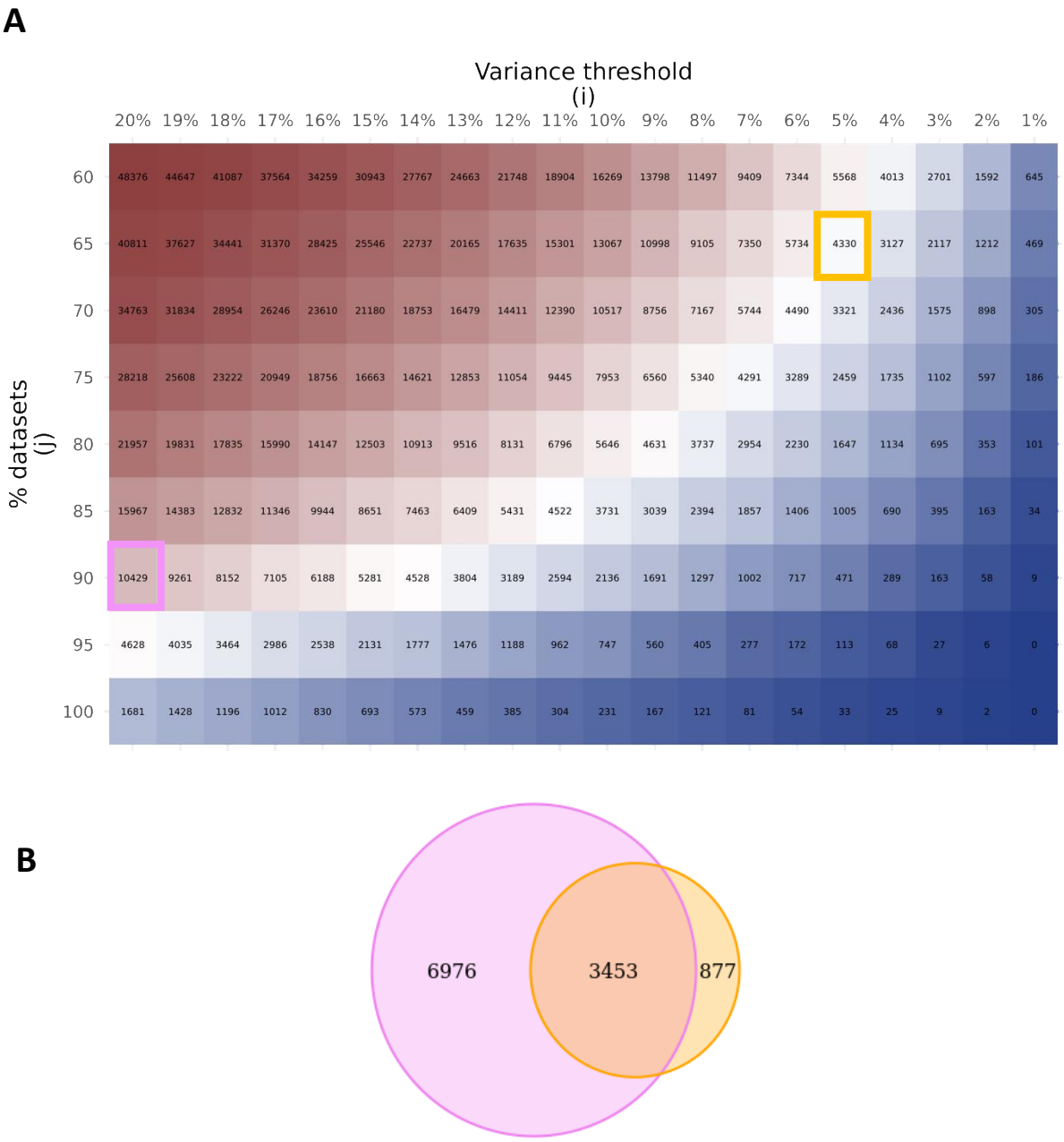

Figure S1. **A large proportion of hvCpGs identified using *i,j* = [5,65] overlap an alternative set obtained using *i,j* = [20,90]. A)** Heatmap showing the number of CpGs identified at *i,j* = [5,65] (orange) and *i,j* =[20,90] (purple). **B)** Venn diagram showing the overlap between these two CpG sets.

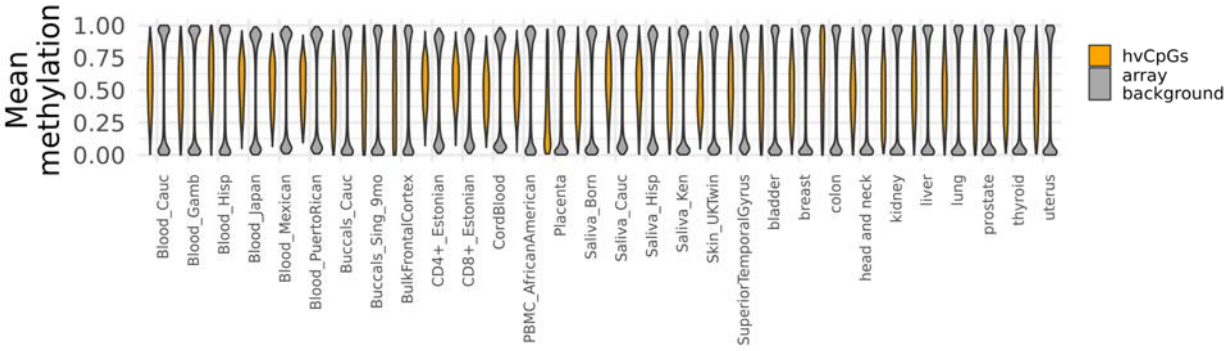

Figure S2. hvCpGs are enriched for intermediate methylation values. The distribution of mean methylation Beta values at hvCpGs (orange) and array background CpGs (grey).

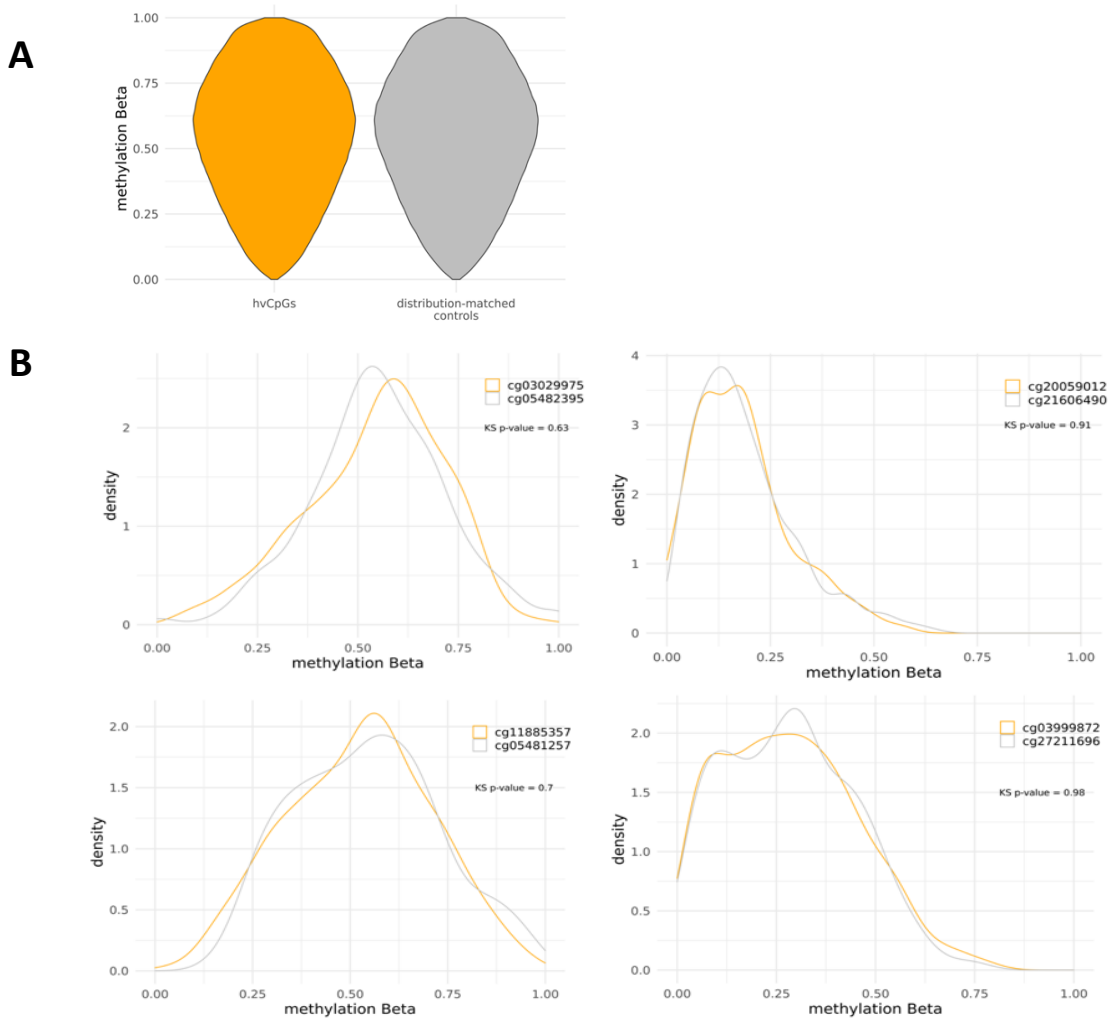

Figure S3. Matching each hvCpG to a control with a similar distribution of methylation Beta values in Caucasian blood. A) The distribution of methylation Beta values at 3,566 hvCpGs and their corresponding distribution-matched controls selected from the 'Blood_Cauc' dataset (Supplementary Table 1). **B)** Examples of the similar distribution of methylation Beta values at hvCpGs (orange) and their corresponding controls (grey). The median Kolmogorov-Smirnov p-value over all 3,566 hvCpG-control pairs was 0.91 (IQR = [0.63, 0.98]). See 'Methods' for further details.

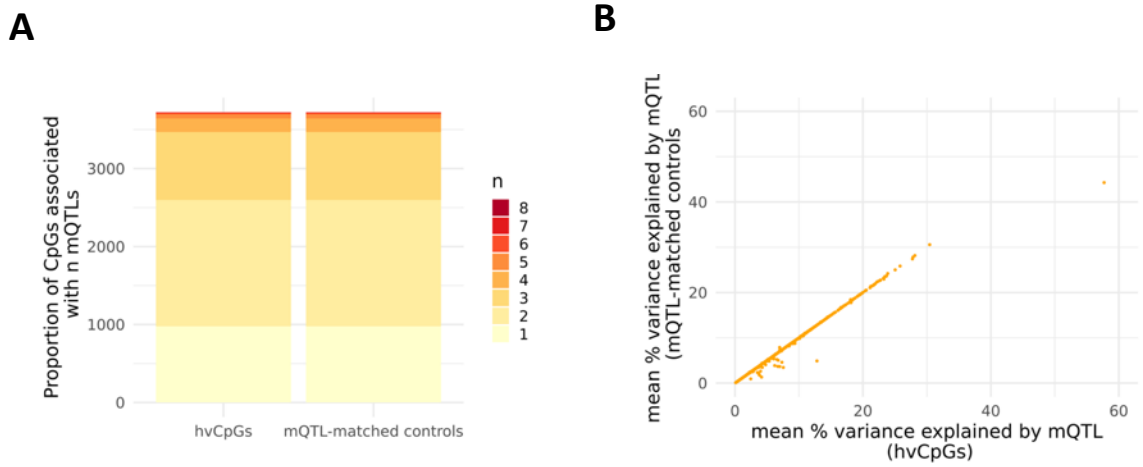

Figure S4. Matching each hvCpG to a CpG with similar mQTL effects using the GoDMC mQTL analysis. Each of the 3,722 hvCpGs reported in the GoDMC mQTL analysis (Min et al. 2021), was matched to a control CpG with **A)** the same number of mQTL associations, and **B)** similar mean % variance explained by mQTL, requiring each control CpG to be present in at least as many datasets as the hvCpG.

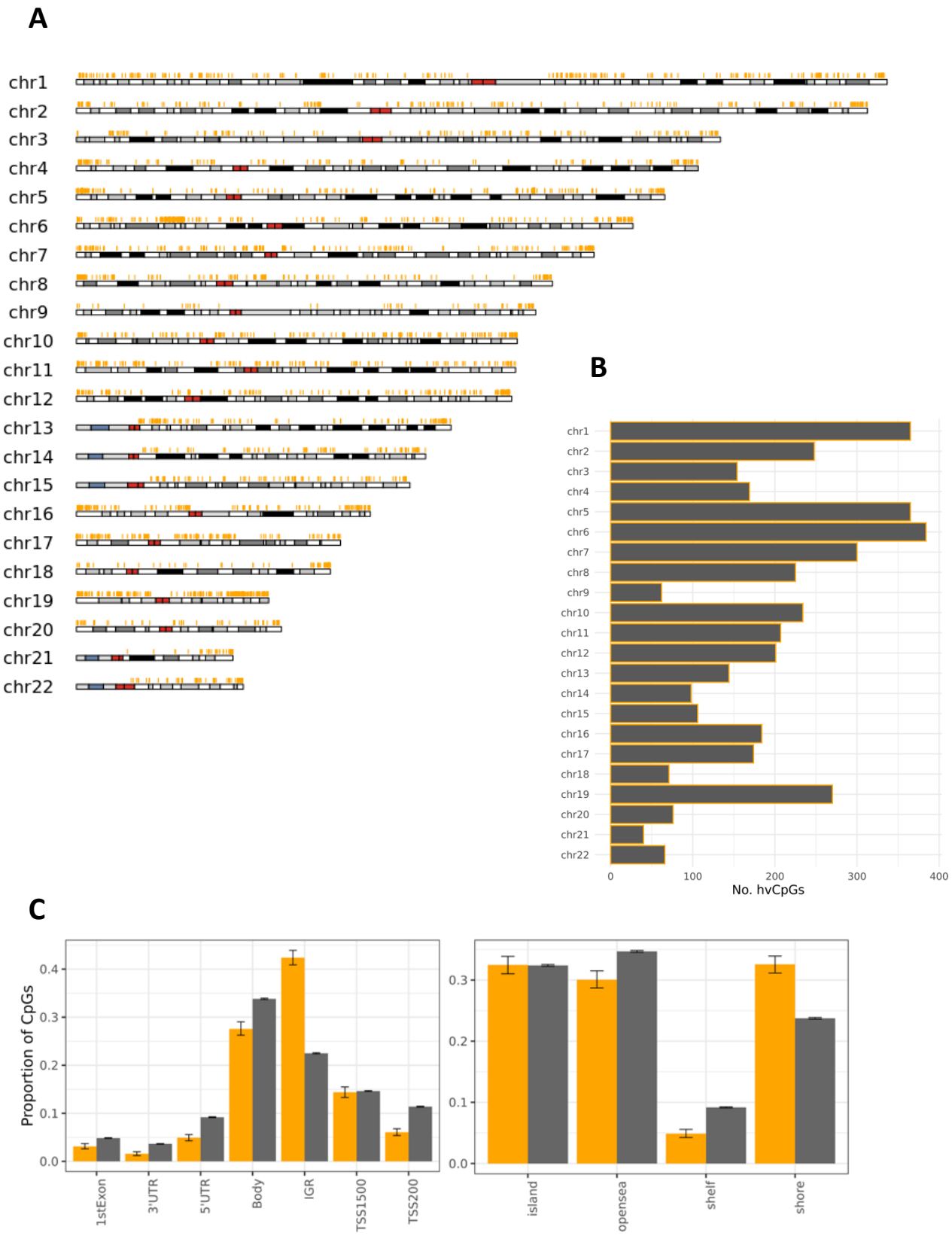

Figure S5. Genomic locations of hvCpGs. A) The positions of hvCpGs across autosomal hg19 chromosomes. **B)** The number of hvCpGs (orange) located on each chromosome. **C).** Genomic locations of hvCpGs (orange) and array background (grey) in relation to gene bodies (left) and CpG islands (right). UTR = untranslated region, IGR = intergenic region, TSS = transcription start site. Error bars are bootstrapped 95% confidence intervals.

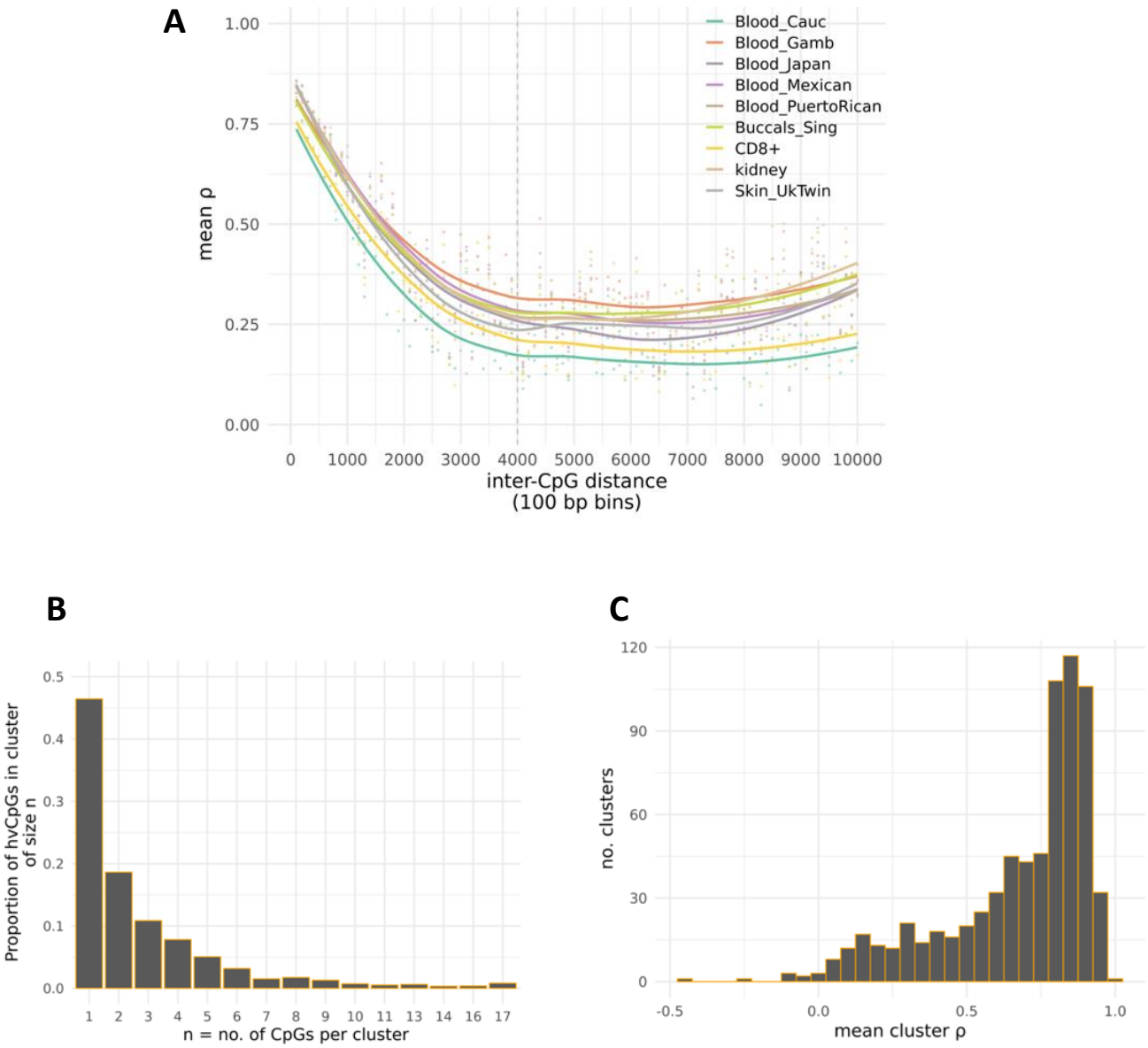

Figure S6. Clustering of hvCpGs. A) The decay of methylation correlation with distance at hvCpGs. Each point indicates the average pairwise Spearman rho across hvCpG pairs with inter-CpG distance falling within each 100 bp bin. Curves are loess lines of best fit for each dataset. Dashed vertical line at 4000 bp is the threshold we used for defining hvCpG clusters. **B)** The proportion of hvCpGs that fall into clusters comprising *n* CpGs, using an inter-CpG distance of 4000 bp to define clusters. **C**) The average Spearman rho across the 716 hvCpG clusters that comprise at least 2 CpGs. For each cluster, we calculated the average pairwise Spearman rho across hvCpG pairs for every dataset in which all CpGs in the cluster were covered, before taking the mean across these datasets. Note, in A and C we only considered datasets with N $\geq$ 100.

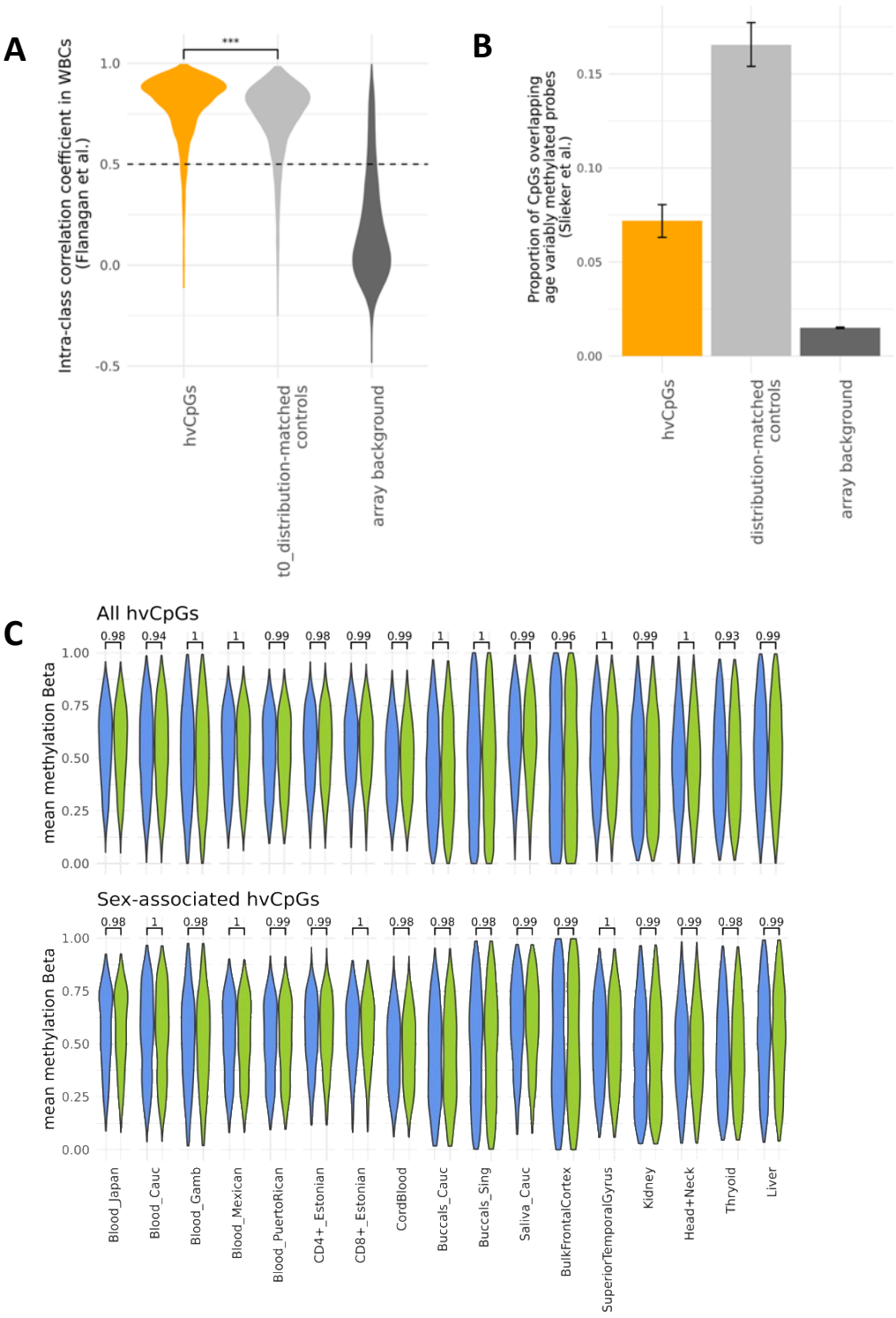

Figure S7. Methylation at hvCpGs is stable with age and is not influenced by sex. A) The distribution of temporal ICC scores at 3,564 hvCpGs and CpGs with similar distribution of methylation Beta values at the first time-point (t0) covered in the Flanagan et al. (2015) dataset (see Methods). The horizontal line at ICC = 0.5 indicates the threshold specified by Flanagan et al. (2015) above which probes are considered to be temporally stable. *** indicates Wilcoxon paired signed-rank test p-value < 0.001. **B)** The proportion of 3,566 hvCpGs and corresponding 'distribution-matched controls' (Table 1) covered in the 'Blood_Cauc' dataset that overlap probes showing increased variability with age in whole blood in a study of 18-to-88-year olds by Slieker et al. (2016). **C)** Comparison of the distribution of mean methylation Beta values between male (blue) and female (green) samples at all hvCpGs (top panel) and at 941 hvCpGs associated with sex in the EWAS catalog (bottom panel) for datasets with N ≥ 50**.** Wilcox Rank Sum p-values are given above each pair of violin plots.

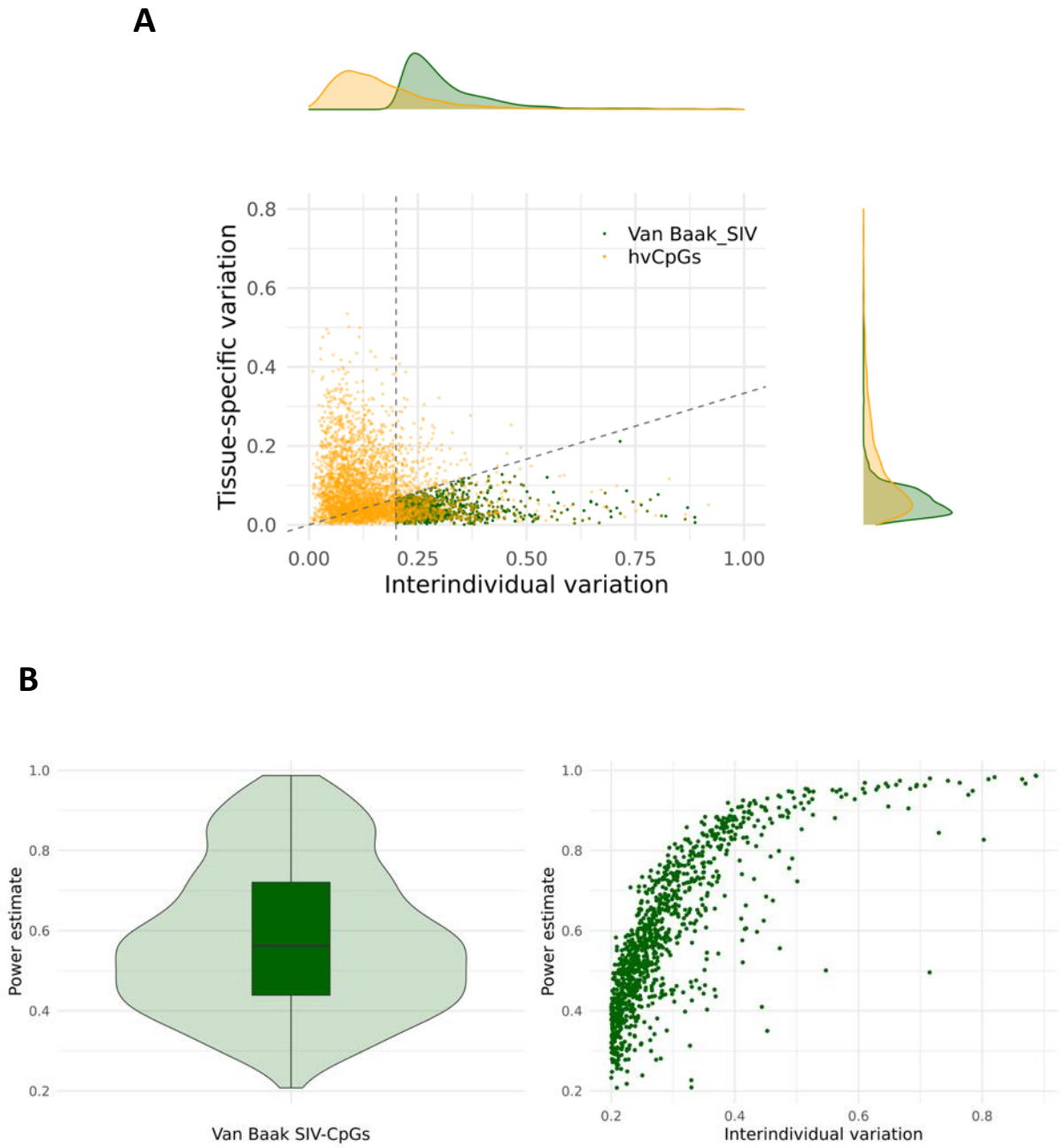

Figure S8. Estimating the power to detect SIV-CpGs reported by van Baak et al. (2018) using a 4-individual multi-tissue dataset (GSE50192). A) Replication of the screen by van Baak et al., in which 1,042 SIV-CpGs (green points) were identified, using the same methylation data from abdominal aorta, gall bladder and ischiatic nerve in 4 individuals. Orange points are hvCpGs. **B**) Estimating the power to detect each of the 1,042 SIV-CpGs reported by van Baak et al. (2018). See Methods for further details**. Left:** distribution of power estimates for each SIV-CpG across 1000 permutations**. Right:** confirmation that estimated power to detect SIV is correlated with interindividual variation.

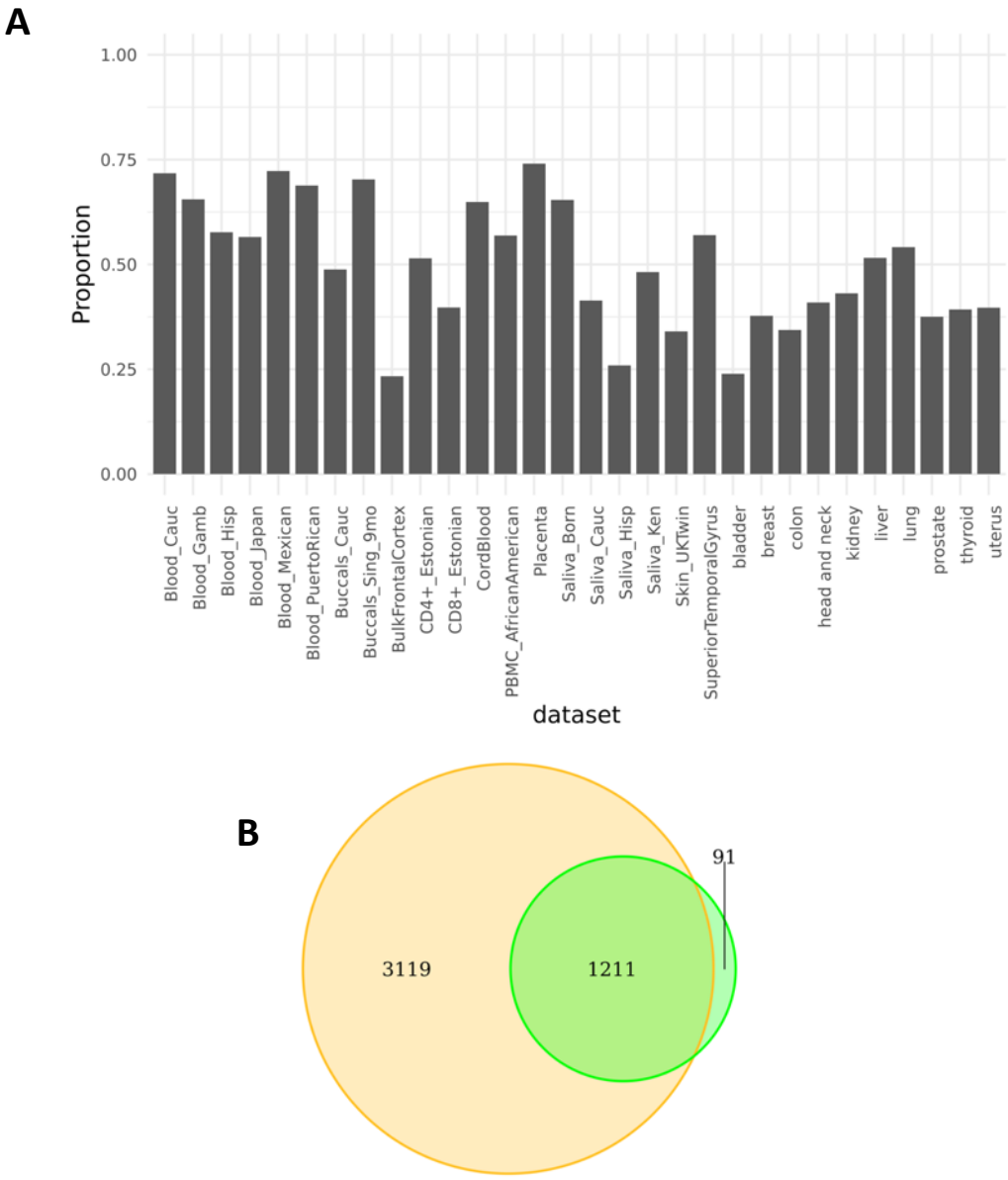

Figure S9. Adjustment for the first 10 PCs of methylation variation in each dataset increases the number of hvCpGs. A) For each dataset, we extracted the top 5% of CpGs by methylation Beta variance before and after adjusting methylation for the first 10 PCs. We then determined the proportionate overlap between PC-adjusted and PC-unadjusted variable CpGs. As expected, PC-adjustment influences the most variable CpGs within each dataset. **B)** The overlap between the set of 4,330 hvCpGs identified using PC-adjusted datasets (orange) and a set of 1,302 hvCpGs identified using unadjusted datasets (green). PC-adjusted: $lm(meth \sim 10\ PCs + age + sex)$, unadjusted: $lm(meth \sim age + sex)$.

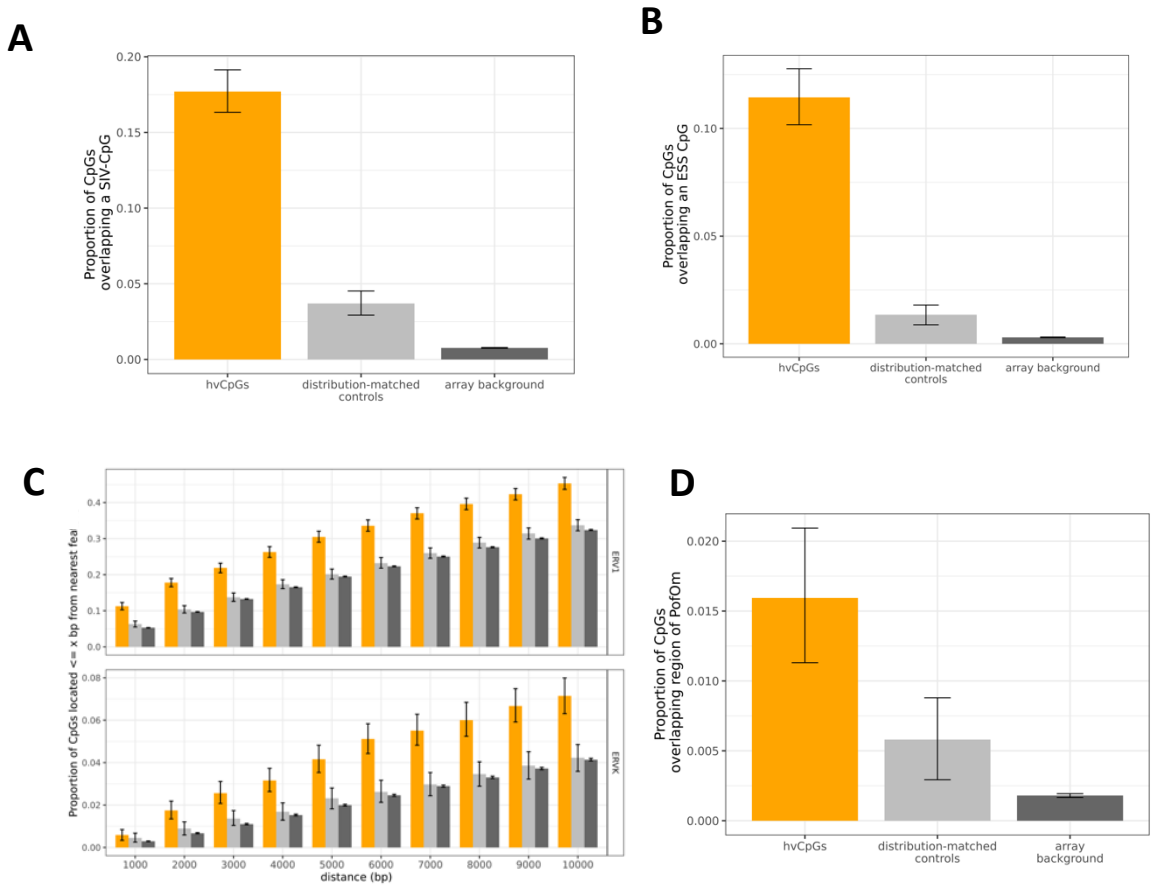

Figure S10. The association between hvCpGs and features suggestive of methylation establishment in the early embryo is maintained when using a de-clustered set of hvCpGs. The proportion of 2,178 de-clustered hvCpGs (in which no CpG is within 4 kb from another) and corresponding distribution-matched controls covered in the 'Blood_Cauc' dataset that **A)** overlap SIV-CpGs (4.8-fold enrichment relative to controls)**, B)** overlap ESS-CpGs (8.5-fold enrichment)**, C)** are located $\leq x$ bp from nearest ERV1 and ERVK transposable elements, and **D)** overlap regions of parent-of-origin specific methylation (PofOm) identified by Zink et al. (2018) (2.7-fold enrichment). See Supplementary Table 12 for further details.

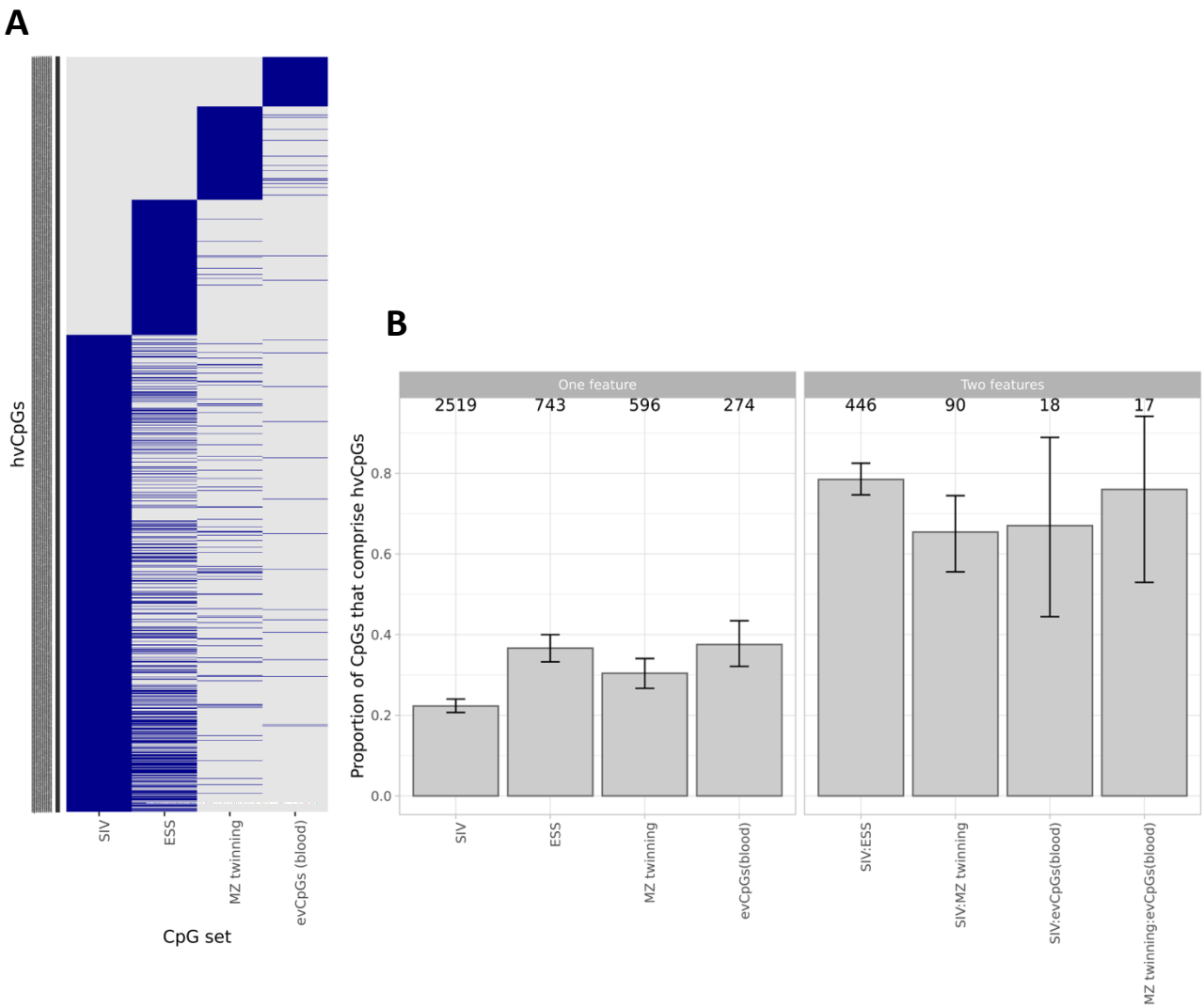

Figure S11. hvCpGs comprise a large proportion of loci that have been previously found to show more than one feature linked to variable methylation establishment in the early embryo. **A)** Heatmap showing the overlap of 1,574 (38%) hvCpGs with SIV, ESS, MZ-twinning CpGs and evCpGs. **B)** The proportion of previously identified CpGs showing one feature (left) or two features (right) linked to the early embryo that comprise hvCpGs. Number of CpGs overlapping the array background is given above each bar, requiring a minimum of 15 CpGs. Error bars are bootstrapped confidence intervals. SIV = systemic interindividual variation, ESS = epigenetic supersimilarity, evCpGs = equivalently variable CpGs.

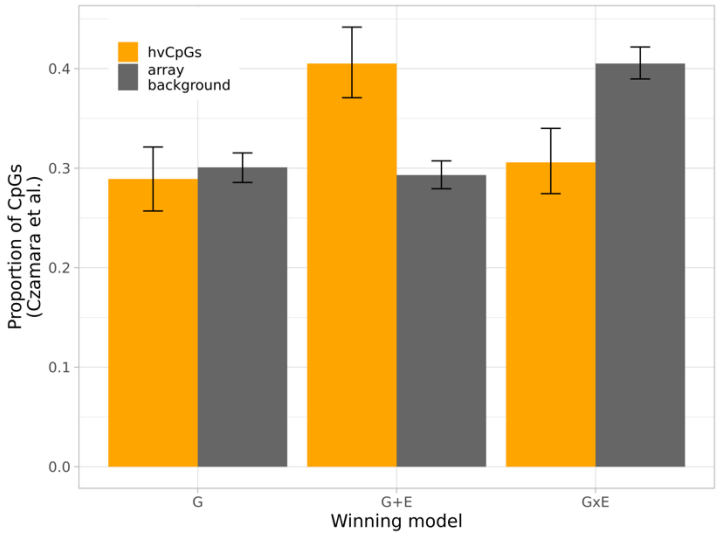

Figure S12. Methylation variability of hvCpGs in whole blood is best explained by G+E effects estimated by Czamara et al. (2019). The proportion of 747 hvCpGs and 3,644 array background CpGs covered by the variably methylated probes (VMPs) reported by Czamara et al. that are best explained by G, G + E and G x E effects. Error bars are bootstrapped 95% confidence intervals.

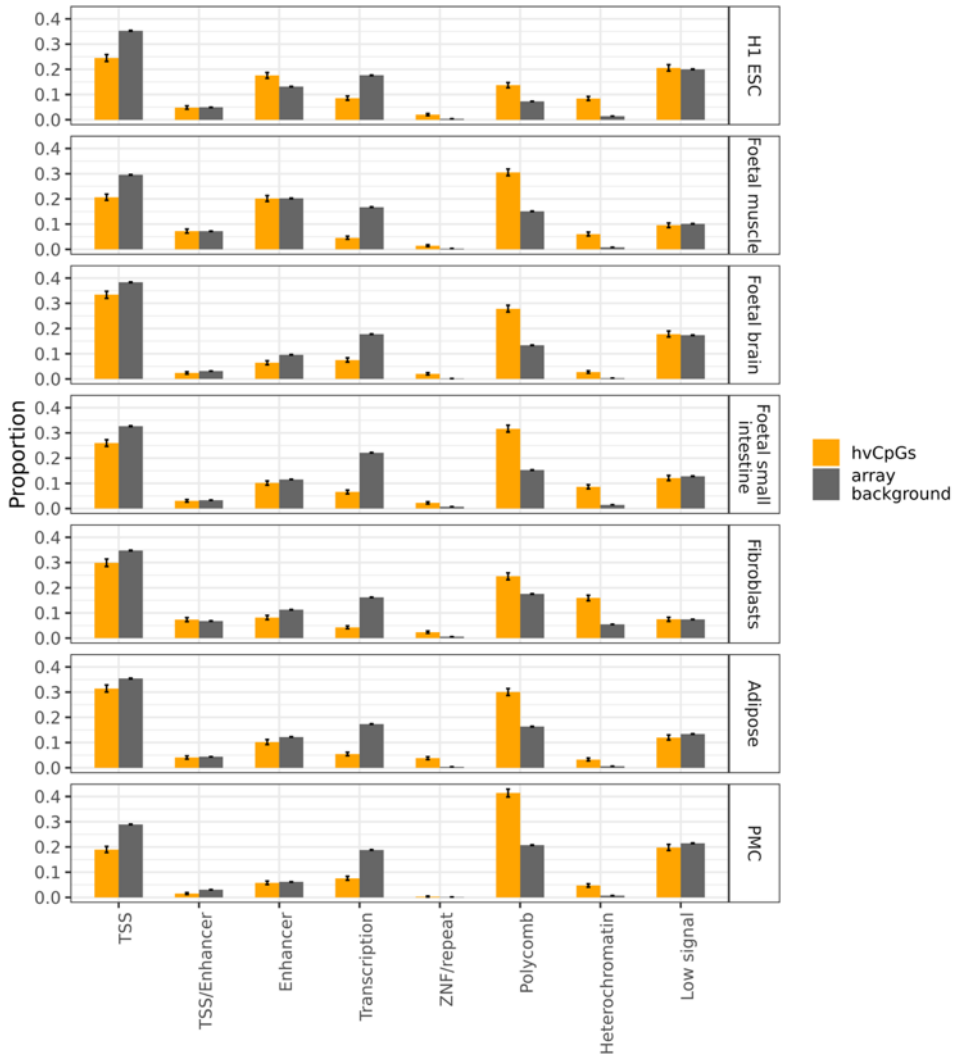

Figure S13. Functional annotation of hvCpGs using the ChromHMM model. The ChromHMM 15-state model using data from H1-hESC, foetal muscle, foetal brain, foetal small intestine, fibroblasts, adipose and primary mononucleocytes from the Roadmap Epigenomics Consortium. 15-states have been collapsed to 8 states for clarity (Supplementary Table 13). TSS = transcription start site. ZNF = zinc finger gene. Error bars indicate bootstrapped 95% confidence intervals.

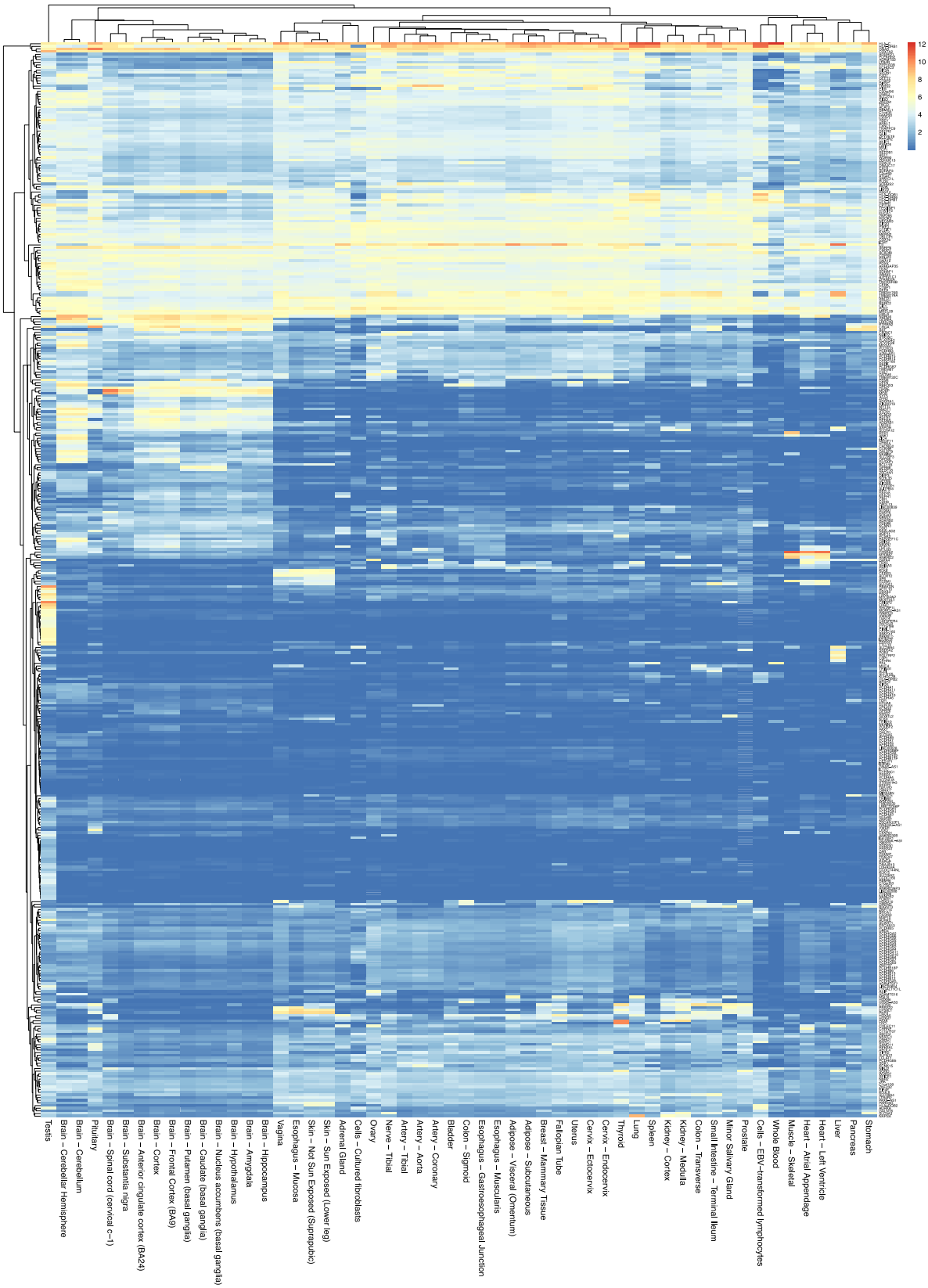

Figure S14. GTEx RNA expression of hvCpG-associated genes. Shown are the 416 genes that are annotated to hvCpG clusters (2 or more hvCpGs no more than 4kb apart). TPM = transcripts per million.