Environmentally sensitive hotspots in the methylome of the early human embryo

Matt J. Silver[1*], Ayden Saffari[1], Noah J. Kessler[2], Giriraj R. Chandak[3], Caroline H.D. Fall[4], Prachand Issarapu[3], Akshay Dedaniya[3], Modupeh Betts[1], Sophie E. Moore[1,5], Philip T. James[1], David Monk[6,7] and Andrew M. Prentice[1]

[1]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, UK & Gambia.
[2]Department of Genetics, University of Cambridge, UK.
[3]Genomic Research on Complex Diseases (GRC Group), CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India.
[4]MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, Southampton, UK.
[5]Department of Women and Children's Health, King's College London, London, UK.
[6]Biomedical Research Centre, University of East Anglia, UK.
[7]Bellvitge Institute for Biomedical Research, Spain.
[*]Corresponding author: matt.silver@lshtm.ac.uk

ABSTRACT

In humans, DNA methylation marks inherited from sperm and egg are largely erased immediately following conception, prior to construction of the embryonic methylome. Exploiting a natural experiment of cyclical seasonal variation including changes in diet and nutritional status in rural Gambia, we replicated 125 loci with a common season-of-conception methylation signature in two independent child cohorts, providing evidence of environmental effects on DNA methylation in the early embryo that persist at least until mid-childhood. Bioinformatic analysis revealed that these loci were highly enriched for metastable epialleles, parent-of-origin specific methylation and regions hypomethylated in sperm, and for H3K9me3 and H3K27me3 histone marks in multiple tissues. They tended to co-locate with endogenous retroviral (ERV1, ERVK) elements. Identified loci were influenced but not determined by measured genetic variation, notably through gene-environment interactions. To the extent that early methylation changes impact gene expression, environmental sensitivity during early embryo genomic remethylation could thus constitute a sense-record-adapt mechanism linking early environment to later phenotype.

KEYWORDS: DNA methylation; early embryo; conception; nutrition; metastable epialleles; gene-environment interactions; parent-of-origin effects; H3K9me3; H3K27me3

DNA methylation (DNAm) plays an important role in a diverse range of epigenetic processes in mammals including cell differentiation, X-chromosome inactivation, genomic imprinting and the silencing of transposable elements[1]. DNAm can influence gene expression and can in turn be influenced by molecular processes including differential action of methyltransferases and transcription factor binding[2–4].

The human methylome in extensively remodelled in the very early embryo when parental gametic methylation marks are largely erased before acquisition of lineage and tissue-specific marks at implantation, gastrulation and beyond[5]. The days following conception may therefore offer a window of heightened sensitivity to external environmental influences, potentially stretching back to the period before conception coinciding with late maturation of oocytes and spermatozoa at loci that (partially) evade early embryonic reprogramming[6].

The effects of early exposures on the mammalian methylome have been widely studied in animals but multiple factors make this challenging in humans. Causal pathways are difficult to elucidate in observational studies, and even randomised experimental designs are prone to confounding due to reverse causation from exposure-related postnatal effects[7].

Here we address these limitations by exploiting a natural experiment in rural Gambia where conceptions occur against a background of repeating annual patterns of dry ('harvest') and rainy ('hungry') seasons with accompanying significant changes in energy balance, diet composition, nutrient status and rates of infection[8,9]. We interrogate early embryonic events by leveraging published data on loci with evidence for the establishment of variable methylation states in the early embryo that persist in post-gastrulation and postnatal tissues; namely loci demonstrating systemic interindividual variation (SIV)[10,11] and/or epigenetic supersimilarity (ESS)[11] (Fig. 1). These loci bear the hallmarks of metastable epialleles (MEs), loci with variable methylation states that were first identified in isogenic mice. MEs exhibit stable patterns of SIV indicating stochastic establishment of methylation marks prior to gastrulation when tissue differentiation begins[12], and several MEs have been shown to be sensitive to periconceptional nutrition in mice[13]. These loci thus serve as useful tools for studying the effects of early environment on DNAm by enabling the use of accessible tissues (such as blood) that can serve as a proxy for systemic (cross-tissue) methylation, and by pinpointing the window of exposure to the periconceptional period[14].

In this study we assess the influence of seasonality on DNAm in two Gambian child cohorts[15,16], enabling robust identification of loci showing consistent effects at the ages of 24

months and 8-9 years (Fig. 1). Through prospective study designs, we capture conceptions throughout the year and use statistical models that make no prior assumptions about specific seasonal windows driving DNAm changes in offspring. We probe potential connections between season of conception (SoC)-associated loci and putative MEs and investigate links with transposable elements and transcription factors associated with the establishment of methylation states in the early embryo. We also assess the influence of genetic variation and gene-environment interactions. Finally, by comparing our results with public DNAm data obtained from sperm, oocytes and multi-stage human embryos, we investigate links between SoC-associated loci, histone marks, gametic and parent-of-origin specific methylation, and the establishment of DNAm states in early embryonic development.

The developmental origins of health and disease hypothesis posits the existence of mechanisms linking prenatal nutrition to lifelong metabolic disease[6]. It has also been suggested that epigenetic mechanisms driving phenotypic variation would be advantageous in the face of changing environments, and that for such mechanisms to have evolved, the propensity to vary should be under genetic control[17]. Our description of genetically directed environmentally sensitive hotspots providing a durable record of conditions during gametic maturation and in the very early embryo fulfils both these predictions.

RESULTS

Association of DNA methylation with Gambian season of conception

Key characteristics of the Gambian cohorts and samples analysed in this study are provided in Table 1. DNAm differences associated with season of conception are potentially confounded by season of sample collection effects in the ENID (discovery) cohort since samples are collected at age 2yrs (Fig. 2A top). This is not the case in the EMPHASIS (replication) cohort where all samples are collected in the Gambian dry season (Fig. 2A bottom). To compare year-round DNAm signatures across cohorts we focussed on 391,814 autosomal CpGs ('array background') intersecting the Illumina HM450 and EPIC arrays used to measure DNAm in the discovery and replication cohorts respectively. We modelled the effect of date of conception on DNAm using Fourier regression[18] which makes no prior assumptions about specific seasonal windows driving DNAm changes in offspring (see Methods).

We began by identifying 1,861 loci ('discovery CpGs') showing significant seasonal variation in 2-year olds from the discovery cohort with a false discovery rate (FDR)<10%. We then analysed seasonal effects at these loci in 8-9-year olds from the replication cohort. Fourier regression models revealed a heterogeneous distribution of year-round methylation peaks and nadirs at discovery CpGs in each cohort (Fig. 2B, Supplementary Table 1). Next, we identified a subset of 125 'SoC-CpGs', defined as CpGs from the discovery CpG set with evidence of significant seasonality (FDR<10%) in the replication cohort (Supplementary Table 2).

SoC-CpGs showed a highly consistent seasonal pattern across both cohorts in marked contrast to matched controls with similar methylation distributions (Fig. 2C; Spearman rho=0.56, $p=8.6x10^{-12}$ and rho=0.02, p=0.65 respectively for conception date at methylation maximum; see Table 2 for details of matched controls).

60% of SoC-CpGs exist as singletons, defined as having no SoC-CpG within 1,000bp, and 85% fall within clusters of 3 CpGs or fewer (Supplementary Table 3). SoC-CpGs are distributed throughout the genome (Supplementary Fig. 1) and include several CpG clusters extending over more than 500bp, notably at *IGF1R* which spans 1,323bp and covers 9 CpGs (Supplementary Table 4, Supplementary Fig. 2). Compared to array background, SoC-CpGs are highly enriched for intermediate (25-75%) methylation states, most notably at putative MEs and loci exhibiting SIV/ESS previously identified in multi-tissue screens in adult Caucasians, hereafter named 'MEs' for short (Figure 2D; Supplementary Table 5; see Table 2 for details of ME/SIV/ESS loci). SoC-CpGs are enriched at CpG islands compared to array background and matched controls (Fig. 2E).

SoC-CpGs and non-replicating discovery CpGs show a distinct pattern of methylation maxima for conceptions falling within the July-September period in both cohorts (Fig. 3A). This pattern is particularly marked at SoC-CpGs (Fig. 3A and 3B top), and also at ME loci in the set of non-replicating discovery CpGs (Fig. 3B top). The July-September period corresponds to the peak of the Gambian rainy season, a strong validation of our previous studies in babies and infants that focussed on conceptions at peak seasons only, with similar observations of higher methylation in conceptions at the peak of the Gambian rainy season compared to peak dry season[11,19–21]. Methylation minima fall within the February-April period, corresponding to the peak of the dry season (Supplementary Fig. 3).

Seasonal methylation amplitude, defined as the difference between modelled methylation peak and nadir, is also significantly greater at SoC-CpGs, and at replicating and non-replicating MEs, compared to controls (Figs. 3A and 3B bottom; Supplementary Table 6; Wilcoxon Rank-Sum test p-value ranging from 7.1×10^{-8} to 1.5×10^{-59}). Furthermore, there is evidence of a substantial and significant decrease in seasonal amplitude at non-replicating MEs in the older cohort (Fig. 3B bottom; median amplitude decrease=4.5%; Wilcoxon $p=6.7 \times 10^{-13}$), and a small but significant decrease at SoC-CpGs that are not known MEs (median decrease=1.0%; $p=1.7 \times 10^{-4}$; Supplementary Table 6). There is no corresponding significant amplitude decrease in replicating MEs or controls.

Compared to array background, SoC-CpGs are highly enriched for MEs (18-fold enrichment, $p=6.8 \times 10^{-10}$). ME enrichment is even greater when accounting for CpG clustering (Supplementary Table 7). No enrichment is observed at matched controls.

Finally, pairwise methylation states are highly correlated at a large majority of SoC-CpGs in both cohorts, so that the same individuals tend to have relatively high or low methylation at multiple SoC-CpGs. This is in marked contrast to discovery CpGs and controls (Fig. 3C), and pairwise correlations are not driven by increased correlation within SoC-CpG clusters (Supplementary Fig. 4). As expected, a small number of strong negative pairwise SoC-CpG correlations (Fig. 3C) include CpGs with methylation maxima in dry season conceptions (as shown in Fig. 2C).

Early stage embryo, gametic and parent-of-origin specific methylation

Given the strong enrichment for MEs within the set of SoC-CpGs, we next analysed links to methylation changes in early stage human embryos, as we have done previously for putative MEs identified in a whole-genome bisulfite-seq (WGBS) multi-tissue screen[10]. We aligned our data with public reduced representation bisulfite-seq (RRBS) data from human IVF embryos[5] and obtained informative methylation calls for 67,870 array background CpGs covered at ≥ 10x read depth in both inner cell mass (ICM, pre-gastrulation) and embryonic liver (post-gastrulation) tissues. We found a highly distinctive pattern of increased intermediate methylation (10-90%) at SoC-CpGs in post-gastrulation embryonic liver tissue. This strongly contrasted with matched controls and with a general trend of genome-wide hyper- and hypo-methylation at loci mapping to array background (Fig. 4A). We observed a similar pattern at

MEs irrespective of their association with SoC, confirming our previous observation[10], here in the subset of MEs present on Illumina arrays (Fig. 4A, 'all MEs').

We previously observed consistent hypomethylation at MEs across all gametic and early embryonic developmental stages, most notably in sperm[10]. We tested the latter observation at SoC-CpGs by aligning our data with public sperm WGBS data[22], restricting our analysis to 292,240 CpGs mapping to array background that were covered at ≥ 10x. 94 SoC-CpGs were covered in the WGBS dataset and these showed a marked decrease in predicted sperm methylation, with 82% [95% CI: 74-90%] hypomethylated (methylation <10%) in sperm, compared to 43% [38%-47%] and 48% [48-48%] at loci mapping to matched control CpGs and array background respectively (Fig. 4B). Strong enrichment for sperm hypomethylation was also observed at ME CpGs (Fig. 4B, 'all MEs'). Postnatal intermediate methylation states at SoC-CpGs were preserved in both Gambian cohorts irrespective of predicted sperm methylation states, in contrast to array background CpGs where methylation distributions strongly reflected predicted sperm hypomethylation status (Fig. 4C left).

Our observation of increased sperm hypomethylation at SoC-associated loci, together with existing evidence that imprinted genes may be especially sensitive to prenatal exposures[21,23,24], prompted us to investigate a potential link between SoC-sensitivity and parent-of-origin specific methylation (PofOm). A recent study used phased WGBS methylomes to identify regions of PofOm in 200 Icelanders[25]. We analysed 699 of these PofOm CpGs overlapping Illumina array background (Table 2) and observed very strong enrichment for PofOm CpGs at SoC-CpGs and at all MEs on the array (44- and 15-fold enrichment, $p=2.2\times10^{-12}$ and 1.8×10^{-36} respectively; Supplementary Table 8). PofOm enrichment at SoC-CpGs is driven by a large (8 CpG) PofOm region at *IGF1R*, and a single replicating ME-SoC-CpG proximal to the human imprinted 14q32 region (Supplementary Table 2). This is reflected in substantially reduced enrichment after adjustment for CpG clustering (Supplementary Table 8). Strong enrichment for PofOm at MEs is maintained after adjustment for clustering.

Regions of PofOm detected in postnatal samples tend to be differentially methylated in gametes[25], and may thus have evaded the widespread epigenetic reprogramming that occurs in the pre-implantation embryo[26]. We tested this directly by interrogating data from a whole-genome screen for germline differentially methylated regions (gDMRs) that persist to the blastocyst stage and beyond[27]. In this analysis, gDMRs were defined as contiguous 25-CpG

regions that were hypomethylated (mean DNAm < 25%) in one gamete and hypermethylated (mean DNAm > 75%) in the other. We began by observing very strong enrichment for oocyte (maternally methylated), but not sperm gDMRs, at all PofOm loci identified by Zink *et al*[25] (Supplementary Table 8), confirming previous observations of an excess of PofOm loci that are methylated in oocytes only[25]. We found a particularly strong 122-fold enrichment for oocyte gDMRs (oo-gDMRs) persisting in placenta. We next analysed SoC-CpGs and MEs and again found evidence for strong enrichment of oocyte, but not sperm gDMRs, at these loci (6.5-fold oo-gDMR enrichment, $p=3.5x10^{-9}$ at SoC-CpGs; 2.9-fold, $p=1.2x10^{-24}$ at MEs; Supplementary Table 8). Once again enrichment at SoC-CpGs and MEs was particularly marked at oo-gDMRs persisting in placenta and was partially driven by CpG clustering (Supplementary Table 8). Of note, more than 4 times as many SoC-CpG clusters were identified as oo-gDMRs persisting in placenta (9 clusters), as were recorded as exhibiting PofOm identified by Zink *et al* (2 clusters; Supplementary Table 2). Note that a large majority of SoC-CpGs that are hypomethylated in sperm are not oo-gDMRs (i.e. they are not hypermethylated in oocytes) (Fig. 4C, bottom right), suggesting that factors associated with regional sperm hypomethylation rather than differential gametic methylation may be a key driver of sensitivity to periconceptional environment at these loci.

Enrichment of DNase I hypersensitive sites and histone marks

We assessed overlap of SoC-CpGs and controls with putative tissue-specific functional elements using eFORGE v2.0[28,29]. As with previous analyses, we adjusted for CpG clustering using a proximity cutoff of 1kbp. We began by analysing enrichment for DNase I hypersensitive sites (DHS) identified in 39 different tissues and cells from the Roadmap Epigenomics Consortium[28]. We found no evidence of significant enrichment of DHS at SoC-CpGs or matched or random control CpGs. We next looked for enrichment of 5 distinct histone (H3) marks identified in the same Roadmap tissues and cells. We found evidence of extremely strong enrichment at SoC-CpGs for H3K9me3 and H3K27me3 derived from multiple sources including fetal and postnatal primary tissues and cells (Fig. 4D). This was in marked contrast to H3 enrichment at matched control CpGs which showed extremely strong enrichment for H3K4me1 and H3K27me3 (Supplementary Fig. 5), consistent with the identification of enhancers associated with non-allelic intermediate methylation profiles[30]. No enrichment for H3 marks was observed at random controls.

Enrichment of transposable elements and transcription factors associated with genomic imprinting

Variable methylation states at MEs are associated with transposable elements (TEs) in murine models[31,32], and we have previously observed enrichment for specific proximal endogenous retroviruses (ERVs) at putative human MEs[10,21]. Here we found some evidence for enrichment of proximity to ERV1 and ERVK at SoC-CpGs, although enrichment was reduced after accounting for CpG clustering (Supplementary Table 7).

Enrichment for PofOm and gDMRs at SoC-CpGs loci suggests a potential link to mechanisms implicated in the maintenance of PofOm and genomic imprinting in the early embryo. Our previous analysis of MEs identified from WGBS data found enrichment for proximal binding sites for 3 transcription factors (TFs: CTCF, ZFP57 and TRIM28) identified through ChIP-seq of embryonic stem and kidney cells that are linked to maintenance of PofOm at imprints[10]. Here we found no evidence for enrichment of proximal bindings sites for these TFs at SoC-CpGs on the Illumina array (Supplementary Table 7).

Influence of genotype and gene-environment interactions

Genetic variation, primarily in *cis*, is a major driver of inter-individual variation in DNAm[33]. A previous analysis quantified methylation variance explained by additive genetic effects, and common and non-shared environment in 1,464 twin pairs from the British E-Risk study[34]. We began by reproducing the result from Hannon *et al*[34] that non-shared environment (which includes measurement error) explains the major part of methylation variance in array background (Fig. 5A). Methylation variance explained by additive genetic effects at SoC-CpGs was markedly higher in comparison (Fig. 5A). On the assumption that SoC-CpGs will be enriched for MEs exhibiting systemic (cross-tissue) interindividual variation (SIV), this supports the finding by Hannon *et al* of increased heritability at loci that are more correlated in blood and brain, suggestive of SIV[34].

We next explored the influence of genotype and environment on DNAm directly in the EMPHASIS (replication) cohort, for which we had genotype data on 288 individuals measured at >3M SNPs after imputation from the Illumina Global Screening Array (GSA; see Methods). For this analysis we assessed genome-wide SNP-DNAm associations to identify methylation quantitative trait loci (mQTL), and also looked for gene-environment (SoC) interactions (GxE)

on DNAm, following a similar strategy to that used in a recent study in the GUSTO cohort[35] (see Methods). Since there is limited power to detect genetic effects in a sample of this size, particularly when assessing interactions, we compared our findings at SoC-CpGs with matched and random control CpGs (see Table 2).

We began by selecting the most significant ('winning') mQTL (G1) and GxE (G2) SNP for each SoC-CpG (Supplementary Table 9). The vast majority of G1 and G2 SNPs mapped to a single SoC-CpG, with the exception of two G1 SNPs each mapping to 2 CpGs in *trans*; and two G2 SNPs, one mapping to 2 CpGs and the other to 8 CpGs in the *IGF1R* SoC-CpG cluster, both also in *trans* (Supplementary Table 10). Furthermore, no major G1 or G2 SNP clusters were discernible (Supplementary Fig. 6. and Supplementary Tables 11 and 12). Together these observations suggest that DNAm differences at SoC-CpGs are not driven by genetic variants at any specific locus covered by the imputed GSA data.

To assess the potential for genetic confounding of SoC-associated DNAm signals, we tested each winning G1 and G2 SNP for association with SoC using 5 different genotypic models. No SNPs passed a Bonferroni-adjusted significance threshold accounting for the number of SNPs and models tested (Supplementary Table 13), thus providing no strong evidence of genetically driven confounding of SoC-associated DNAm.

We next ran a series of Fourier regression models to determine the relative proportions of methylation variance explained by E (periconceptional environment only), G1 (mQTL only) and G2xE (including E and G2 main effects but excluding G1 main effects; see Methods). Results for SoC-CpGs were compared to matched and random control CpGs. Variance explained by E, G1 and G2xE models was assessed using adjusted R^2 values to account for increasing model complexity. In each case adjusted R^2 values were compared to a baseline model that included the same set of covariates (principal components, age and sex) used in Fourier regression models for the main seasonality analysis. At SoC-CpGs, mQTL (G1) models explained significantly more methylation variance than seasonality alone (E models). However, gene-environment (G2xE) models explained significantly more methylation variance than both G1 and E models (Fig. 5B, Supplementary Table 14). A formal assessment of 'winning models', using the Akaike Information Criterion (AIC) to account for differences in model complexity determined that G2xE models provided the best fit for 83% of SoC-CpGs, compared with 24% and 32% for random and high variance controls respectively (Fig. 5B inset).

As expected, year-round DNAm at a CpG where G1 is the winning model indicates a strong mQTL effect on mean methylation (Fig. 5C, bottom left). In contrast, at a CpG where GxE effects dominate, the strength of the seasonality effect is modified by genotype (Fig. 5C, top left; dashed lines); revealing a large seasonal effect that is less apparent when modelling data unstratified by genotype (same figure, solid red line). Scatter plots of underlying individual-level DNAm data adjusted solely for baseline covariates support these observations (Fig. 5C right).

A recent analysis of GxE effects in the GUSTO cohort revealed a similar dominance of GxE effects at a subset of variable CpGs when considering a range of *in utero* environmental effects including maternal BMI, smoking and maternal depression[35]. Speculating that these loci may be similarly sensitive to periconceptional environment, we tested SoC-CpGs and controls for enrichment of 889 GxE CpGs identified by Teh *et al*[35] that overlapped array background. We observed a highly significant 18-fold enrichment of these GxE CpGs amongst SoC-CpGs ($p=1.1x10^{-5}$; cluster-adjusted 26-fold, $p=2.4x10^{-6}$; Supplementary Table 7).

DISCUSSION

We have exploited a natural, seasonal experiment in rural Gambia whereby human conceptions are 'randomised' to contrasting environmental (especially dietary) conditions to examine whether these differential exposures leave a discernible signature on the offspring methylome. We identified 125 'SoC-CpGs' with strong evidence of sensitivity to season of conception in independent, different-aged cohorts. Importantly, these cohorts have contrasting confounding structures, notably with regard to the timing of sample collection; the latter eliminating potential confounding due to seasonal differences in leukocyte composition. These results, derived from analysis of Illumina array data with limited coverage, suggest there may be many more hotspots sensitive to the periconceptional environment across the human methylome.

This analysis builds on previous epigenetic studies in this setting that have focussed on single cohorts and analysed methylation differences between individuals conceived at the peaks of the Gambian dry and rainy seasons only[11,19–21,36]. Greater methylation in offspring conceived at the peak of the Gambian rainy season is consistent with previous findings and this observation is now greatly strengthened by the application of Fourier regression to model year-round conceptions – an approach that makes no prior assumption of when methylation

peaks and nadirs may occur. The number of identified SoC-CpGs is also substantially increased in this study. Comparisons with array background and control CpGs matching SoC-CpG methylation distributions, together with a two-step discovery-replication design, increase confidence that these findings are not statistical artefacts.

Multiple lines of evidence support the notion that methylation states at these loci are established in the early embryo. First, they are highly enriched for putative MEs and related loci identified in other studies with characteristic methylation signatures suggestive of establishment early in embryonic development[10,11]. Second, like MEs, season-associated loci exhibit highly unusual methylation dynamics in early stage embryos[10]. Third, they have distinctive gametic methylation patterns, notably hypomethylation in sperm (in common with putative MEs[10]), and differential gametic and parent-of-origin specific methylation in a subset. Increased sperm hypomethylation at SoC-CpGs may reflect their enrichment at CpG islands[37] (Fig. 2E), sequence features that are largely refractory to protamine exchange, with the possibility for retaining epigenetic function associated with histone modifications into the early embryo[38]. Fourth, they are highly enriched for H3K27me3 and H3K9me3 marks which, along with the TRIM28/KAP1-repressor complex, play pivotal roles in restraining transposon expression whilst protecting imprinted PofOm in pre-implantation embryos undergoing reprogramming[39]. Furthermore, these heterochromatic marks also coordinate transient gene expression and early lineage commitment, in part through demarking active and poised enhancers that are frequently associated with intermediate methylated states[30,40].

A large majority of SoC-CpGs have not previously been identified as MEs, but given the supporting evidence described above, we speculate that many are likely to be so. Indeed, evidence of attenuation of SoC effects with age suggests that, to the extent that interindividual variation is driven by periconceptional environmental factors, screens for putative MEs (including ESS and SIV) in adult tissues used as a reference in this analysis may be missing metastable regions which are more pronounced earlier in the life course. SoC effect attenuation in the older cohort could also explain the lack of replication of SoC associations at the majority of CpGs from the discovery set, despite suggestive evidence of a common effect of periconceptional environment (Figs. 3A and 3B). Importantly, this would have implications for detecting the effect of periconceptional exposures on DNAm in samples collected beyond the neonatal and early childhood periods, an important consideration for epigenetic epidemiological studies since non-persisting methylation differences could still have a significant impact on early developmental trajectories with life-long consequences[41,42].

Methylation states at SoC-associated loci are highly correlated within individuals despite being distributed throughout the genome, strongly suggesting that a common mechanism is at play. This contrasts with a recent study of murine MEs located within intracisternal A particle insertions (IAPs, of which the *Agouti* locus is a paradigm example[43]), where no intra-individual correlation between stochastic methylation states was observed, although the mice were not exposed to different environments[32].

Potential insights into mechanisms linking periconceptional environment to DNAm changes in postnatal tissues come from our investigations of the methylation status and genomic context of SoC-CpGs.

First, we observed a strong overlap of SoC-CpGs with regions that are known to be hypomethylated in sperm. A minority of these loci are hypermethylated in oocytes with PofOm persisting in postnatal tissues. This latter observation aligns with a growing body of evidence linking early environment, notably nutritional factors involved in one-carbon (C1) metabolism, with methylation at imprinted regions[23,24]. Indeed we have previously noted an association between season of conception and several C1 metabolites at a maternally imprinted region at the small non-coding RNA *VTRNA2-1*[21], consistent with evidence of 'polymorphic imprinting' linked to prenatal environment at this locus[25,44]. Furthermore, we previously found strong enrichment for proximal binding sites of several transcription factors (TFs) associated with the maintenance of PofOm in the early embryo at MEs detected in a WGBS screen[10]. We were unable to replicate this at SoC-CpGs in this study which may reflect the relatively small proportion of PofOm loci in the set of SoC-CpGs, or factors related to the biased methylome coverage of Illumina arrays. More targeted experimental work is required to determine the extent of SoC effects at imprinted loci, especially given our observation that SoC-CpGs are enriched for heterochromatic histone modifications and proximity to ERV transposable elements that have recently been shown to drive the establishment of germline-derived maternal PofOm[45].

Second, we observed modest enrichment for proximity to ERV1 and ERVK transposable elements at SoC-CpGs. This was also observed at MEs on the Illumina array, confirming our previous observations[10,21]. Enrichment of ERVs at SoC-CpGs is notable since most environment-sensitive mouse MEs are associated with IAPs (which are rodent-specific ERVs)[32], and Krab zinc-finger protein (KZFP)-mediated repression of transposable elements (TEs) including ERVs has been proposed as a driver of the rapid evolution of gene regulation[46].

The KZFP ZFP57 is particularly interesting in this respect since its binding to DNA is linked both to repression of TEs and to the maintenance of genomic imprints in the pre-implantation embryo[23,47]. We previously identified a putative SoC-associated DMR in the *ZFP57* promoter in blood from Gambian infants[21], and a proximal CpG 21kbp from this DMR is in the set of discovery CpGs in this study, indicating a putative SoC association in Gambian 2 year-olds. It is possible that non-replication of the SoC-association at *ZFP57* in the older Gambian cohort reflects the more general attenuation of SoC effects described above. Interestingly there is some evidence that the *ZFP57* DMR, which lies 3kb upstream of the transcription start site, is established in the early embryo[11]. Given the important function of ZFP57 in pre-implantation methylation dynamics, its potential role as an environmentally sensitive regulator of genome-wide SoC effects on DNA remains an open question.

Third, DNAm at SoC-CpGs is highly enriched for intermediate methylation states. Intermediate methylation has also been observed at MEs in Gambians and in non-Africans[19,20,48–50], and this coincides with a similar observation at MEs in post-gastrulation embryonic tissues[10]. This latter observation includes measurements from single conceptuses, with methyl-seq read-level analyses indicating that intermediate methylation is driven by adjacent CpGs which all tend to have the same methylation state on a given DNA molecule [10].

Taken together, the above evidence suggests that a periconceptional environmental exposure may perturb methylation by nudging the ratio of methylated to unmethylated DNA molecules at hotspots in the early post-gastrulation embryo. These hotspots appear to be concentrated in regions that are hypomethylated in sperm, and, in the case of PofOm, additionally hypermethylated in oocytes. In the latter case, methylation states could be driven by an environmentally sensitive gain of methylation on the paternal allele that is propagated through development, incomplete reprogramming on the maternal allele leaving residual traces, or modest *de novo* methylation at some later point. A deeper understanding of mechanisms will require further investigation in cell and animal models.

Several SoC-CpGs with evidence of PofOm map to an intronic region of the *IGF1R* gene. Zink *et al*[25] were unable to demonstrate PofO allele-specific expression (PofO-ASE) in this region although others have found evidence of maternal imprinting of an intronic lncRNA at this gene in cancerous cells[51,52]. Loss of IGF1 receptors gives rise to a major decrease in expression at multiple imprinted genes in mice, suggesting a pathway by which *IGF1R* might regulate growth and metabolism during early development[53]. IGF1R signalling is implicated in fetal

growth, glucose metabolism and cancer[54–56], and DNAm differences at *IGF1R* have been observed in birthweight-discordant adult twins[57]. Another SoC-associated locus with PofOm is approximately 300kbp from the 14q32 *DLK1-MEG3* imprinted region, close to the imprinted C14MC microRNA cluster[58], and within 80kb of a region with PofO-ASE[25]. Epigenetic and transcriptional changes at several C14MC microRNAs have been implicated in cancer[59–61], and genetic and epigenetic mutations in the 14q32 region are linked to imprinting disorders including Temple syndrome[62,63].

Another notable SoC-CpG is within 1000bp of a metastable variably methylated region (VMR) at the intron2/exon3 boundary of the *POMC* gene. *POMC* is a key regulator of appetite through the production of melanocyte-stimulating hormones in the hypothalamus[50]. There is evidence that hypermethylation at the VMR reduces *POMC* expression by interfering with P300 TF binding at the intron2/exon3 boundary of the gene[64], and is linked to the presence of a primate-specific *Alu* element (transposon)[65]. This region has previously been associated with SoC and certain C1 metabolites in Gambian infants[50], and is associated with obesity in German children and adults[50,64]. It is interesting to note that hypermethylation of the *POMC* SoC-CpG and the VMR occurs in conceptions at the height of the Gambian rainy season, a period also known as the 'hungry season' when food stocks from the previous year's harvest are depleted. A link between *POMC* VMR hypermethylation established in the early embryo that persists into postnatal life, reduced *POMC* expression and corresponding reduced satiety signalling, could therefore constitute a 'predictive-adaptive-response', whereby an individual's early developmental trajectory is tuned to its anticipated postnatal environment[66].

DNAm is influenced by genotype and the latter is therefore a potential confounder when studying the effects of environmental exposures in human populations. A strength of our quasi-randomised Gambian seasonal model is that it minimises the potential for genetic confounding of modelled seasonal DNAm patterns, on the assumption that the timing of conceptions is not linked to genetic variants influencing DNAm. However, it is still possible that such variants might confound our observations, for example if they promote embryo survival under conditions of environmental stress. We tested this possibility using genetic data available for the EMPHASIS (replication) cohort, and found no evidence of SoC-associated genetic variants driving inter-individual methylation differences at SoC-associated loci in *cis* or *trans*.

We did however uncover evidence of gene–periconceptional environment interactions at SoC-CpGs that explained a greater proportion of methylation variance than environmental or direct genetic factors alone. While our analysis had limited power, confidence in this finding is increased through our comparison of genetic effects at SoC-CpGs with matched and random controls. The relatively strong influence of GxE effects at SoC-CpGs was supported by very strong enrichment for CpGs showing gene–*in utero* environment interaction effects that similarly explained a greater proportion of methylation variance in a study of the Singaporean GUSTO cohort[35]. Widespread GxE interaction effects could manifest through the action of environmental factors on gene variant-associated transcription factors, although we found no evidence of clustered genetic variants driving these effects at multiple SoC-CpGs. In the context of this study, widespread GxE interaction effects on DNAm would lead to reduced power to detect SoC associations, suggesting that these associations will be easier to detect in adequately powered analyses stratified by genotype.

We have previously argued that the definition of MEs should be extended to include genomic regions whose DNAm state is under partial but non-deterministic genetic influence in genetically heterogeneous human populations[10], and we would argue that the above observations at SoC-CpGs that exhibit many of the characteristics of MEs support this. Further analysis in larger datasets with genome sequencing data combined with functional analysis using cell models will be required to fully understand the relative contributions of environment and genetics to DNAm variation at regions of the type highlighted in this study.

There is increasing interest in the phenomenon of methylation variability as a marker of disease and of prenatal adversity[67,68], and in genetic variation as a potential driver of methylation variance[69]. An intriguing possibility suggested by our gene-environment interaction analysis is that certain genetic variants could have been selected through their ability to enable graded, environmentally-responsive methylation patterns at MEs and SoC-associated loci that are able to sense the periconceptional environment, record the information, and adapt the phenotype accordingly. This mechanism was previously proposed in a theoretical population genetic model of selectable phenotypic variation in changing environments[17]. As discussed with reference to periconceptional programming of the *POMC* gene above, such a mechanism would be adaptive where phenotypic development is directed to better fit the anticipated future environment, but may otherwise become maladaptive, leading to later disease, if the environment changes[6]. Further work to identify signatures of selection at genetic variants driving gene-environment interactions will be required before

we can determine the extent to which environmentally sensitive hotspots might act in this way.

METHODS

Gambian cohorts and sample processing

Detailed descriptions of the Gambian cohorts analysed in the season of conception study are published elsewhere[15,16]. Briefly, for the younger cohort, blood samples from 233 children aged 2 years (median[IQR]: 731[729,733] days old) were collected from participants in the **E**arly **N**utrition and **I**mmune **D**evelopment ("ENID") study[15]. DNA was extracted, bisulfite-converted and hybridised to Illumina HumanMethylation450 (hereafter "HM450") arrays following standard protocols (see Van Baak *et al*[11] for further details). For the older cohort, DNA was extracted from blood samples from 289 Gambian children aged 8-9 (9.0 [8.6,9.2] years) participating in the **E**pigenetic **M**echanisms linking **P**re-conceptional nutrition and **H**ealth **As**sessed in **I**ndia and **S**ub-Saharan Africa ("EMPHASIS") study[16], and was bisulfite-converted and hybridised to Illumina Infinium Methylation EPIC (hereafter "EPIC") arrays, again using standard protocols.

For the ENID cohort, date of conception was calculated from fetal gestational age estimates obtained by ultrasound at the mother's first 'booking' appointment. The same method was used for the EMPHASIS cohort, except for n=71 pregnancies that were > 24 weeks gestation at booking meaning that GA could not be accurately determined by ultrasound[16,70]. In this case date of conception was calculated as date of birth minus 280 days which is the average gestational length for this population.

Methylation array pre-processing and normalisation

Raw intensity IDAT files from the HM450 and EPIC arrays were processed using the *meffil*[71] package in R (v3.6.1) using standard *meffil* defaults. Briefly, this comprised probe and sample quality control steps (filtering on bisulfite conversion efficiency, low probe detection p-values and bead numbers, high number of failed samples per probe, high number of failed probes per sample); methylation-derived sex checks; removal of ambiguously mapping (i.e. cross-hybridising) probes; removal of probes containing SNPs at the CpG site or at a single base extension; and removal of non-autosomal CpGs. Following filtering, methylation data was

normalised with dye-bias and background correction using the *noob* method[72], followed by Functional Normalisation to reduce technical variation based on principal component analysis of control probes on the arrays[73]. After pre-processing and normalisation, methylation data comprised methylation Beta values for 421,026 CpGs on the HM450 array for 233 individuals from the ENID cohort, and 802,283 CpGs on the EPIC array for 289 individuals from the EMPHASIS cohort. Finally 391,814 CpGs intersecting both arrays were carried forward for statistical analysis.

Statistical modelling

Variation of DNAm with date of conception was modelled using Fourier regression[18,74]. This models the relationship between a response variable (here DNAm) and a cyclical predictor (date of conception). The effect of the latter is assumed to be cyclical due to annually varying seasonality patterns, so that the modelled effect for an individual conceived on the 31st December should be 'close' to that for an individual conceived on the 1st of January. This is achieved by deconvolving the conception date (predictor) into a series of pairs of sin and cosine terms, and obtaining estimates for the regression coefficients β and γ in the following model:

$$M_{ij} = \alpha_{0j} + \sum_{k=1}^{m} \alpha_{ik} + \sum_{r=1}^{n} [\beta_{rj}\sin(r\theta_i) + \gamma_{rj}\cos(r\theta_i)] + \varepsilon_{ij}$$

Where, for individual i and CpG j:

M_{ij} is the logit-transformed methylation Beta value[75];

α_{0j} is an intercept term;

α_{ik} is the kth of m adjustment covariates;

θ_i is the date of conception in radians in the interval [0, 2π], with 1st January = 0 and 31st December = 2π, modelled as n pairs of Fourier terms, $\sin\theta_i + \cos\theta_i + ... + \sin n\theta_i + \cos n\theta_i$;

β_r and γ_r are the estimated regression coefficients for the rth sin and cosine term respectively;

and ε_{ij} is the error term.

With a single pair of Fourier terms (n=1), this gives a sinusoidal pattern of variation, with a single maximum and minimum whose phase (position in the cycle) and amplitude (distance

between maximum and minimum) is determined by β_1 and γ_1, with the constraint that the maximum and minimum are 6 months apart. More complex patterns of seasonal variation are afforded by higher frequency pairs of Fourier terms (r>1).

For this analysis we modelled the effect of date of conception using a single pair of Fourier terms and assessed goodness of fit by comparing full and covariate-only models using likelihood ratio tests. For both cohorts, covariates included child sex, and the first six principal components (PCs) obtained from unsupervised principal component analysis (PCA) of the normalised methylation M-values. The PCs were used to account for unmeasured and measured technical variation (due to bisulfite conversion sample plate, array slide etc) and cell composition effects (see Supplementary Tables 15 and 16). Additional checks confirmed no seasonal variation in estimated white cell composition in either cohort (see below). 450k Sentrix Column was included as an additional adjustment covariate for the ENID cohort since this was not robustly captured by any of the first 6 PCs (Supplementary Table 15). Child age was included as an additional adjustment covariate for the EMPHASIS cohort, since child ages ranged from 8 to 9 years, as was maternal nutritional intervention group (see Chandak *et al*[16] for further details).

For CpG j, coefficient estimates β_j, γ_j were determined by fitting a model with a single pair of Fourier terms (n=1) using *lm()* in R. Model goodness-of-fit was determined by likelihood ratio test (LRT) using *lrtest*() in R, comparing the full model including Fourier terms, with a baseline covariates-only model. A model p-value, p_j was then derived from the corresponding LRT chi-squared statistic. Thus for a given threshold, α, $p_j<\alpha$ supports rejection of the null hypothesis that for CpG j, the full model including the effect of seasonality modelled by one pair of Fourier terms, fits no better than the covariate-only model at the α level.

Identification of 'discovery CpGs' and 'SoC-CpGs'

For the discovery (ENID) cohort, CpG p-values, p_j, were used to compute a false discovery rate for each CpG accounting for multiple testing (assuming 391,814 independent tests corresponding to the number of loci in array background) using *p.adjust()* in R. 1,861 CpGs with a FDR<10% formed the set of discovery CpGs.

In the replication analysis, CpGs from the discovery CpG set were analysed in the replication (EMPHASIS) cohort using the same regression modelling approach. Assuming 1,861

independent tests, 125 CpGs had a FDR<10% and these formed the set of replicating 'SoC-CpGs'.

Additional modelling of seasonal variation in blood cell composition

Cell count estimates using the Houseman method[76] were obtained using the *estimateCellCounts*() from *minfi* (v1.30.0) in R. Seasonal variation in blood cell composition was then modelled by Fourier regression with one pair of Fourier terms and sex (ENID+EMPHASIS) and age (EMPHASIS only) as adjustment covariates. Fitted models indicated no consistent or marked seasonal differences within and between cohorts (Supplementary Figure 7).

CpG sets considered in analyses

Summary information on curated sets of CpGs considered in the analyses is provided in Table 2. Further information on these is provided below.

i. *1,881 ME* CpGs overlap one or more of the following curated sets of loci: putative MEs exhibiting SIV identified in a multi-tissue WGBS screen in Kessler *et al*[10]; and CpGs exhibiting 'epigenetic supersimilarity' and/or SIV described in Van Baak *et al*[11].
ii. *699 parent-of-origin-specific CpGs (PofOm CpGs)* overlapping 229 regions with PofOm identified in Supplementary Table 1 from Zink et al[25].
iii. 889 GxE CpGs listed in Supplementary Table 6 from Teh et al[35]. These are highly variable loci where methylation variance is best explained by GxE models, with E covering a range of *in utero* exposures.

Selection of control CpGs

5 '**matched control**' CpGs drawn from array background (excluding discovery CpGs and known MEs/ESS/SIV CpGs) were identified for each of the 125 SoC-CpG, making a total of 625 CpGs. Matched controls were selected to have similar methylation Beta distributions to SoC-CpGs using a two-sided Kolmogorov-Smirnov test for divergence of cumulative distribution functions (*ks.test()* in R) with a p-value threshold p>0.1. Examples are given in Supplementary Fig. 8, along with a comparison of sample mean distributions.

An additional 625 **random control** CpGs were randomly sampled from array background, again excluding discovery CpGs and known MEs/ESS/SIV CpGs.

Early stage embryo and sperm methylation data

RRBS methylation data from Guo *et al*[5] was downloaded from GEO (accession number GSE49828). Only CpGs covered at ≥ 10x in pre-gastrulation inner cell mass and post-gastrulation embryonic liver were considered in this analysis. Further details are provided in Kessler *et al*[10].

Sperm methylation data from Okae *et al*[22] was downloaded from the Japanese Genotype-phenotype Archive (accession number S00000000006). Only CpGs covered at ≥ 10x were considered in this analysis.

Germline gDMRs

Germline DMRs (gDMRs), defined as contiguous 25 CpG regions that were hypomethylated (DNAm mean +1SD < 25%) in one gamete and hypermethylated (DNAm mean-1SD > 75%) in the other, were previously identified by Sanchez-Delgado *et al*[27]. Persistence of PofOm to the blastocyst and placental stages was established by identifying overlapping intermediately methylated regions in the relevant embryonic tissues, with confirmation of PofOm expression at multiple DMRs[27]. See Sanchez-Delgado *et al*[27] for further details.

Transposable elements and transcription factors

Transposable element (TE) regions determined by RepeatMasker were downloaded from the UCSC hg19 annotations repository, since the reference coordinates for the HM450 and EPIC chip probes, as well as the genomic data from previous analyses and other public datasets used in this study used this reference. Further details on TE regions and ZFP57, TRIM28 and CTCF transcription factor binding sites identified from ChIP-seq in human embryonic kidney and human embryonic stem cells used in this analysis are described in Kessler *et al*[10].

Enrichment tests

Enrichment for overlap of 'test set' CpGs (SoC-CpGs, matched controls etc) with 'feature set' CpGs (MEs, Teh *et al* GxE CpGs, CTCF, ZFP57, TRIM28 TF binding sites and ERVs) was determined by Fisher Exact Tests, with the enrichment odds ratio (OR) defined as

$$\mathrm{OR} = \frac{\text{n test CpGs overlapping feature CpG / n test CpGs not overlapping feature CpG}}{\text{n non-test CpGs overlapping feature CpG / n non-test CpGs not overlapping feature CpG}}$$

MEs, PofOm, gDMR and GxE features were tested for direct overlap with test set CpGs. TF binding sites and ERVs were tested for overlap within 10,000bp for comparison with results from Kessler *et al*[10].

Cluster-adjusted enrichment tests were performed on 'de-clustered' test sets, obtained as follows:

i. Create CpG clusters formed from adjacent CpGs where each CpG is within 1,000bp of the nearest neighbouring CpG;
ii. Construct de-clustered test set by randomly sampling a single CpG from each cluster; non-clustered 'singleton' CpGs are always selected.

In the case of SoC-CpGs, the set of 125 non-clustered CpGs were reduced to 91 CpGs after de-clustering.

Tests for enrichment of DNase I hypersensitive (DHS) sites and histone (H3) marks were performed by uploading CpG sets to the *eForge2.0* website (https://eforge.altiusinstitute.org/). Tests used the following options: proximity filter - 1kb; platform - Illumina 450k (corresponding to array background); background repetitions – 1,000. Datasets used for enrichment tests: DHS enrichment – 'Consolidated Roadmap Epigenomics - DHS'; H3 enrichment - 'Consolidated Roadmap Epigenomics – H3.

Genetic association analyses

<u>Genotype data</u>

Gene-DNAm association analyses were performed on all 288 individuals from the EMPHASIS (replication) cohort for which we had QC'd genotype data. 125 SoC-CpGs, plus sets of 625 matched and random control CpGs (defined above) were considered in this analysis. Subjects were genotyped using the Illumina Infinium Global Screening Array-24 v1.0 Beadchip

(Illumina, California, U.S.) following standard protocols[77]. Array-derived genotypes were pre-phased using SHAPEITv2[78] and imputation was performed using IMPUTEv2.3.2 on 1000 genome phase 3 data[79]. SNPs with a MAF $\leq$ 10% were excluded, along with those with an IMPUTE 'info' metric $\leq$ 0.9, a stringent threshold to ensure maximum confidence in imputation quality. Finally, to minimise the influence of low frequency homozygous variants in linear models, analysis was restricted to SNPs with 10 or more homozygous variants, resulting in a final dataset comprising 3,124,453 SNPs.

Identification of 'winning' mQTL (G1) and gene-environment interaction (G2) SNPs

Environment (E), genome-wide mQTL (G) and gene-environment (GxE) associations were assessed using the *GEM* package (v1.10.0) from R Bioconductor[80], following a similar strategy to that described in Teh *et al*[35].

3 separate models were considered for each CpG, *j:*

1. *E-model:*

   M_j ~ covs + sinθ + cosθ

   This is the same model used in the main Fourier regression analysis for the EMPHASIS cohort described above, with seasonality modelled as one pair of Fourier terms and covs corresponding to the same adjustment covariates used in the main analysis.

2. *G-model:*

   M_j ~ covs + G

   Here, G is SNP genotype coded as allelic dosage (0,1,2) and covs are adjustment covariates as described above.

3.

4. *GxE-model*

   M_j ~ covs + sinθ + G + G x sinθ

   when sinθ is the most significant Fourier term in E-model

   OR

   M_j ~ covs + cosθ + G + G x cosθ

   when cosθ is the most significant Fourier term in E-model.

   Here, G and covs are as described above.

For each CpG the winning 'G1' and 'G2' SNPs were selected as the SNP with the smallest p-value for G and GxE model coefficients respectively. Models with winning G1 and G2 SNPs are referred to as G1 and G2xE models below.

E, G1 and G2xE model comparisons

To account for model complexity (i.e. differing numbers of terms in regression models), comparisons of methylation variance explained by E, G1 and G2xE models (Figure 5b bar plots, Supplementary Table 14) are based on adjusted R-squared values. In each case, for each CpG

$$\Delta adjR^2 = adjR^2_{model} - adjR^2_{cov}$$

where $adjR^2_{model}$ is the adjusted R^2 value for the E, G1 and G2xE models (as described above), and $adjR^2_{cov}$ is the adjusted R^2 for the covariate-only model.

Winning models (Figure 5b pie charts) are those with the lowest value of the Akaike Information Criterion[81] (AIC).

G1 and G2xE fitted model plots

G1 and G2xE fitted model plots (Fig. 5C top and bottom left) for each CpG, *j*, are generated from fitted values estimated from the following models:

G1 model fitted values: $M_j \sim covs + \sin\theta + \cos\theta + G1_j$

G2xE model fitted values: $M_j \sim covs + \sin\theta + \cos\theta + G2_j + G2_j \times \sin\theta + G2_j \times \cos\theta$

SoC association analysis

Potential confounding of SoC-DNAm signals by SoC-associated genetic variants was assessed by analysing SoC associations with G1 and G2 SNPs using SNPassoc v.19-2 in R[82]. Hardy-Weinberg equilibrium tests were performed to check for irregular allelic distributions, potentially reflecting genotyping or imputation errors. Codominant, dominant, recessive, overdominant and log-additive genetic models were tested, with SoC modelled as a dichotomous variable (dry SoC=Jan-Jul; rainy SoC=Aug-Dec) in each case.

All bootstrapped confidence intervals presented in this paper use 1,000 bootstrap samples.

Ethics approval and consent to participate

Ethics approval for the Gambian ENID and EMPHASIS trials was obtained from the joint Gambia Government/MRC Unit The Gambia's Ethics Committee (ENID: SCC1126v2; EMPHASIS: SCC1441). The ENID study is registered as ISRCTN49285450. The EMPHASIS study is registered as ISRCTN14266771. Signed informed consent for both studies was obtained from parents, and verbal assent was additionally obtained from the older children who participated in the EMPHASIS study.

Data and code availability

ENID 450k methylation data analysed for this study is deposited in GEO (GSE99863). Gambian EMPHASIS EPIC methylation data is available on request and will be made publicly available once results from the main EMPHASIS study have been published. Sources and locations of other publicly available data used in this analysis are described in METHODS. Bespoke code used in the analysis is available at https://github.com/mattjsilver/SoCFourier.

References

1. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nat. Rev. Genet.* **14**, 204–220 (2013).
2. Jeltsch, A. Molecular Enzymology of Mammalian DNA Methyltransferases. in *DNA Methylation: Basic Mechanisms* 203–225 (Springer-Verlag). doi:10.1007/3-540-31390-7_7
3. Feldmann, A. *et al.* Transcription Factor Occupancy Can Mediate Active Turnover of DNA Methylation at Regulatory Regions. *PLoS Genet.* **9**, (2013).
4. Kriaucionis, S. & Klose, R. J. ATACing DNA Methylation during Differentiation. *Mol. Cell* **77**, 1159–1161 (2020).
5. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
6. Fleming, T. P. *et al.* Origins of lifetime health around the time of conception: causes and consequences. *Lancet* **391**, 1842–1852 (2018).
7. Birney, E., Smith, G. D. & Greally, J. M. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. *PLOS Genet.* **12**, e1006105 (2016).
8. Moore, S. E. *et al.* Prenatal or early postnatal events predict infectious deaths in young adulthood in rural Africa. *Int. J. Epidemiol.* **28**, 1088–95 (1999).
9. Dominguez-Salas, P. *et al.* DNA methylation potential: Dietary intake and blood concentrations of one-carbon metabolites and cofactors in rural African women. *Am. J. Clin. Nutr.* **97**, 1217–1227 (2013).
10. Kessler, N. J., Waterland, R. A., Prentice, A. M. & Silver, M. J. Establishment of

environmentally sensitive DNA methylation states in the very early human embryo. *Sci. Adv.* **4**, eaat2624 (2018).
11. Van Baak, T. E. *et al.* Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol.* **19**, 2 (2018).
12. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* **18**, 348–51 (2002).
13. Anderson, O. S., Sant, K. E. & Dolinoy, D. C. Nutrition and epigenetics: an interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. *J. Nutr. Biochem.* **23**, 853–859 (2012).
14. Gunasekara, C. J. & Waterland, R. A. A new era for epigenetic epidemiology. *Epigenomics* **11**, 1647–1649 (2019).
15. Moore, S. E. *et al.* A randomized trial to investigate the effects of pre-natal and infant nutritional supplementation on infant immune development in rural Gambia: the ENID trial: Early Nutrition and Immune Development. *BMC Pregnancy Childbirth* **12**, 107 (2012).
16. Chandak, G. R. *et al.* Protocol for the EMPHASIS study; epigenetic mechanisms linking maternal pre-conceptional nutrition and children's health in India and Sub-Saharan Africa. *BMC Nutr.* **3**, 81 (2017).
17. Feinberg, A. P. & Irizarry, R. A. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. *Proc. Natl. Acad. Sci.* **107**, 1757–1764 (2010).
18. Rayco-Solon, P., Fulford, A. & Prentice AM. Differential effects of seasonality on preterm birth and intrauterine growth. *Am. J. Clin. Nutr.* **81**, 134–139 (2005).
19. Waterland, R. A. *et al.* Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.* **6**, e1001252 (2010).
20. Dominguez-Salas, P. *et al.* Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat. Commun.* **5**, 1–7 (2014).
21. Silver, M. *et al.* Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. *Genome Biol.* **16**, 118 (2015).
22. Okae, H. *et al.* Genome-Wide Analysis of DNA Methylation Dynamics during Early Human Development. *PLoS Genet.* **10**, e1004868 (2014).
23. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting disorders: lessons on how genome, epigenome and environment interact. *Nat. Rev. Genet.* **20**, 235–248 (2019).
24. James, P. *et al.* 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. *Int. J. Epidemiol.* 1–28 (2018). doi:10.1093/ije/dyy153
25. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat. Genet.* **50**, 1542–1552 (2018).
26. Meyenn, F. Von & Reik, W. Forget the Parents : Epigenetic Reprogramming in Human Germ Cells. *Cell* 1248–1251 (2015).
27. Sanchez-Delgado, M. *et al.* Human Oocyte-Derived Methylation Differences Persist in the Placenta Revealing Widespread Transient Imprinting. *PLOS Genet.* **12**, e1006427 (2016).
28. Breeze, C. E. *et al.* eFORGE: A Tool for Identifying Cell Type-Specific Signal in Epigenomic Data. *Cell Rep.* **17**, 2137–2150 (2016).
29. Breeze, C. E. *et al.* eFORGE v2.0: updated analysis of cell type-specific signal in

epigenomic data. *Bioinformatics* **35**, 4767–4769 (2019).
30. Elliott, G. *et al.* Intermediate DNA methylation is a conserved signature of genome regulation. *Nat. Commun.* **6**, 6363 (2015).
31. Waterland, R. A. & Jirtle, R. L. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.* **23**, 5293–300 (2003).
32. Kazachenka, A. *et al.* Identification, Characterization, and Heritability of Murine Metastable Epialleles: Implications for Non-genetic Inheritance. *Cell* 1–13 (2018). doi:10.1016/j.cell.2018.09.043
33. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
34. Hannon, E. *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.* **14**, 1–27 (2018).
35. Teh, A. L. *et al.* The effect of genotype and in utero environment on inter-individual variation in neonate DNA methylomes. *Genome Res.* (2014). doi:10.1101/gr.171439.113
36. Kühnen, P. *et al.* Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity. *Cell Metab.* **24**, 502–509 (2016).
37. Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. *Cell* **146**, 1029–1041 (2011).
38. Hammoud, S. S. *et al.* Distinctive chromatin in human sperm packages genes for embryo development. *Nature* **460**, 473–478 (2009).
39. Walter, M., Teissandier, A., Pérez-Palacios, R. & Bourc'his, D. An epigenetic switch ensures transposon repression upon dynamic loss of DNA methylation in embryonic stem cells. *Elife* **5**, n/a-n/a (2016).
40. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–285 (2011).
41. Vukic, M., Wu, H. & Daxinger, L. Making headway towards understanding how epigenetic mechanisms contribute to early-life effects. *Philos. Trans. R. Soc. B Biol. Sci.* **374**, 20180126 (2019).
42. Simpkin, A. J. *et al.* Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum. Mol. Genet.* **24**, 3752–3763 (2015).
43. Morgan, H. D., Sutherland, H. G. E., Martin, D. I. K. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.* **23**, 314–318 (1999).
44. Carpenter, B. L. *et al.* Mother–child transmission of epigenetic information by tunable polymorphic imprinting. *Proc. Natl. Acad. Sci.* 201815005 (2018). doi:10.1073/pnas.1815005115
45. Bogutz, A. B. *et al.* Evolution of imprinting via lineage-specific insertion of retroviral promoters. *Nat. Commun.* **10**, 1–14 (2019).
46. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).
47. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
48. Finer, S. *et al.* 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 (2016).
49. Clark, J. *et al.* Associations between placental CpG methylation of metastable epialleles and childhood body mass index across ages one, two and ten in the Extremely Low Gestational Age Newborns (ELGAN) cohort. *Epigenetics* **0**,

15592294.2019.1633865 (2019).
50. Kühnen, P. *et al.* Interindividual Variation in DNA Methylation at a Putative POMC Metastable Epiallele Is Associated with Obesity. *Cell Metab.* **24**, 502–509 (2016).
51. Kang, L. *et al.* Aberrant allele-switch imprinting of a novel IGF1R intragenic antisense non-coding RNA in breast cancers. *Eur. J. Cancer* **51**, 260–270 (2015).
52. Sun, J. *et al.* A novel antisense long noncoding RNA within the IGF1Rgene locus is imprinted in hematopoietic malignancies. *Nucleic Acids Res.* **42**, 9588–9601 (2014).
53. Boucher, J. *et al.* Insulin and insulin-like growth factor 1 receptors are required for normal expression of imprinted genes. *Proc. Natl. Acad. Sci.* **111**, 14512–14517 (2014).
54. Randhawa, R. & Cohen, P. The role of the insulin-like growth factor system in prenatal growth. *Mol. Genet. Metab.* **86**, 84–90 (2005).
55. Aguirre, G. A., Ita, J. R., Garza, R. G. & Castilla-Cortazar, I. Insulin-like growth factor-1 deficiency and metabolic syndrome. *J. Transl. Med.* **14**, 1–23 (2016).
56. Larsson, O., Girnita, A. & Girnita, L. Role of insulin-like growth factor I receptor signalling in cancer. *Br. J. Cancer* **92**, 2097–2101 (2005).
57. Tsai, P.-C. *et al.* DNA Methylation Changes in the IGF1R Gene in Birth Weight Discordant Adult Monozygotic Twins. *Twin Res. Hum. Genet.* **18**, 635–646 (2015).
58. Malnou, E. C., Umlauf, D., Mouysset, M. & Cavaillé, J. Imprinted MicroRNA Gene Clusters in the Evolution, Development, and Functions of Mammalian Placenta. *Front. Genet.* **9**, (2019).
59. Nayak, S. *et al.* Novel internal regulators and candidate miRNAs within miR-379/miR-656 miRNA cluster can alter cellular phenotype of human glioblastoma. *Sci. Rep.* **8**, 7673 (2018).
60. Kumar, A. *et al.* Identification of miR-379/miR-656 (C14MC) cluster downregulation and associated epigenetic and transcription regulatory mechanism in oligodendrogliomas. *J. Neurooncol.* **139**, 23–31 (2018).
61. González-Vallinas, M. *et al.* Epigenetically Regulated Chromosome 14q32 miRNA Cluster Induces Metastasis and Predicts Poor Prognosis in Lung Adenocarcinoma Patients. *Mol. Cancer Res.* **16**, 390–402 (2018).
62. Beygo, J. *et al.* New insights into the imprinted MEG8-DMR in 14q32 and clinical and molecular description of novel patients with Temple syndrome. *Eur. J. Hum. Genet.* **25**, 935–945 (2017).
63. Eggermann, T. *et al.* Imprinting disorders: a group of congenital disorders with overlapping patterns of molecular changes affecting imprinted loci. *Clin. Epigenetics* **7**, 123 (2015).
64. Kuehnen, P. *et al.* An Alu Element–Associated Hypermethylation Variant of the POMC Gene Is Associated with Childhood Obesity. *PLoS Genet.* **8**, e1002543 (2012).
65. Kuehnen, P. & Krude, H. Alu elements and human common diseases like obesity. *Mob. Genet. Elements* **2**, 197–201 (2012).
66. Low, F. M., Gluckman, P. D. & Hanson, M. A. Developmental Plasticity, Epigenetics and Human Health. *Evol. Biol.* **39**, 650–665 (2012).
67. Webster, A. P. *et al.* Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. *Genome Med.* **10**, 1–12 (2018).
68. Tobi, E. W. *et al.* Selective Survival of Embryos Can Explain DNA Methylation Signatures of Adverse Prenatal Environments. *Cell Rep.* **25**, 2660-2667.e4 (2018).
69. Ek, W. E. *et al.* Genetic variants influencing phenotypic variance heterogeneity. *Hum. Mol. Genet.* **27**, 799–810 (2018).
70. Owens, S. *et al.* Periconceptional multiple-micronutrient supplementation and

placental function in rural Gambian women: A double-blind, randomized, placebo-controlled trial. *Am. J. Clin. Nutr.* **102**, 1450–1459 (2015).
71. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. *Bioinformatics* **34**, 3983–3989 (2018).
72. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res.* **41**, 1–11 (2013).
73. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 0–42 (2014).
74. Nabwera, H. M., Fulford, A. J., Moore, S. E. & Prentice, A. M. Growth faltering in rural Gambian children after four decades of interventions: a retrospective cohort study. *Lancet Glob. Heal.* **5**, e208–e216 (2017).
75. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
76. Jaffe, A. E. & Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
77. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
78. Delaneau, O., Marchini, J. & Zagury, J.-F. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–81 (2012).
79. van Leeuwen, E. M. *et al.* Population-specific genotype imputations using minimac or IMPUTE2. *Nat. Protoc.* **10**, 1285–1296 (2015).
80. Pan, H., Holbrook, J. D., Karnani, N. & Kwoh, C. K. Gene, Environment and Methylation (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and integrate genotype and interaction between genotype and environment. *BMC Bioinformatics* **17**, 299 (2016).
81. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).
82. González, J. R. *et al.* SNPassoc: An R package to perform whole genome association studies. *Bioinformatics* **23**, 644–645 (2007).

Acknowledgements

We acknowledge the work of Z. Herceg, M. N. Routledge, Y. Y. Gong, and H. Hernandez-Vargas in acquiring the ENID methylation data.

Funding

The Gambian ENID trial was jointly funded by the UK Medical Research Council (MRC) and the Department for International Development (DFID) under the MRC/DFID Concordat agreement (MRC Program MC-A760-5QX00). Methylation analysis of ENID samples was supported by the Bill & Melinda Gates Foundation (grant no: OPP1 066947). The Gambian EMPHASIS study is jointly funded by MRC, DFID and the Department of Biotechnology,

Ministry of Science and Technology, India under the Newton Fund initiative (MRC grant no.: MR/N006208/1 and DBT grant no.: BT/IN/DBT-MRC/DFID/24/GRC/2015–16). Further support for this analysis was provided by MRC Grant MR/M01424X/1. We acknowledge the work of the full EMPHASIS Study Group (www.emphasisstudy.org) in acquiring this data.

Authors' contributions

MJS conceived the study, performed the bulk of the analysis and wrote the original draft. AS, NJK and PTJ performed additional analyses. PI, AD and MB contributed to sample acquisition and/or processing. MJS and AMP obtained the funding. GRC, CHDF, MJS, SEM, DM and AMP established the cohorts analysed in this study and/or provided data used in the analysis. All authors contributed to and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Main Tables

Table 1. Gambian seasonality-methylation analysis: cohort characteristics

Cohort	sample size	Age	% male	tissue	methylation array
ENID (discovery)	233	2y	50.6	peripheral blood	Illumina Infinium HM450
EMPHASIS (replication)	289	8-9y	54.3	peripheral blood	Illumina Infinium MethylationEPIC

Table 2. CpG sets considered in this analysis. See Methods for further details.

CpG set	Number of CpGs	Notes
Array background	391,814	Intersection of CpGs on Illumina HM450 (discovery) and EPIC (replication) cohort arrays, post QC
Discovery CpGs	1,861	SoC-associated loci identified in the discovery cohort (LRT FDR<10%)
SoC-CpGs	125	Discovery CpGs with significant seasonal variation in the replication cohort (LRT FDR<10%)
Putative metastable epialleles (MEs)	1,881	ME/SIV/ESS CpGs identified in Van Baak *et al*[11] and Kessler *et al*[10] overlapping array background
Parent-of-origin specific methylation (PofOm)	699	Parent-of-origin specific methylation loci identified in Zink *et al*[25] overlapping array background
GxE CpGs	889	CpGs with evidence of G x (*in utero*) E interactions identified in Teh et al.[35] overlapping array background
Matched controls	625	CpGs with similar methylation distributions to SoC-CpGs*
Random controls	625	Random sample of 625 CpGs from array background

QC: quality control; LRT: likelihood ratio test; ESS: epigenetic supersimilarity[11]; G: gene / genetic variant; E: environment / SoC.
*Each SoC-CpG has 5 matched controls, with matching methylation distributions determined by Kolmogorov-Smirnov tests (see Supplementary Fig. 8).

Main Figures

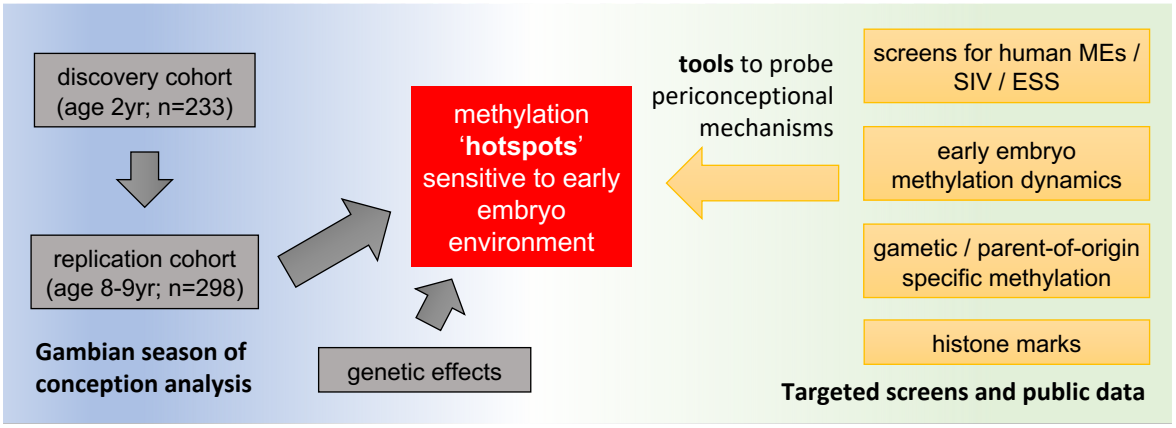


Figure 1. Study design. MEs: metastable epialleles; SIV: systemic interindividual variation; ESS: epigenetic supersimilarity

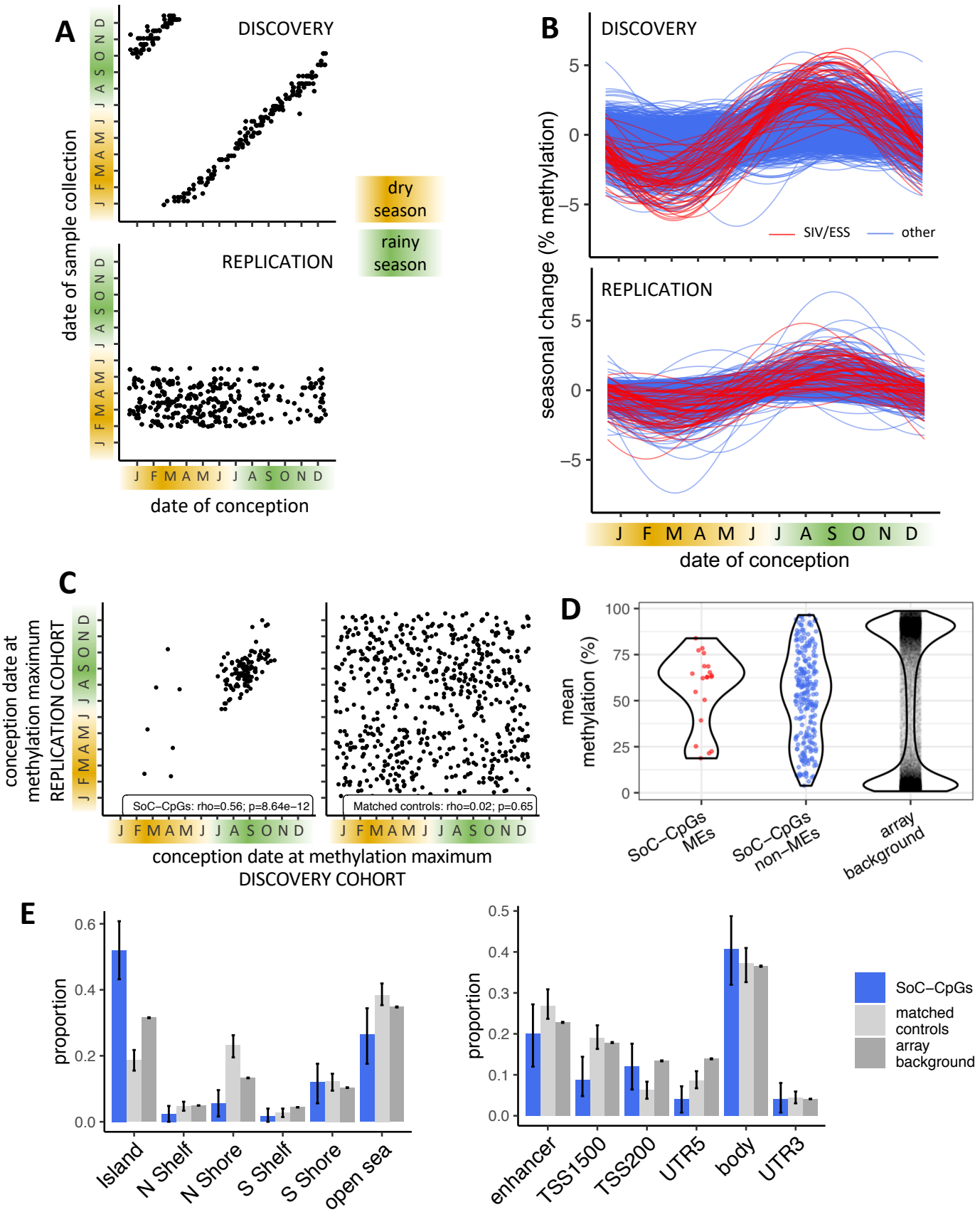


Figure 2. Association of periconceptional environment with DNA methylation in Gambian children. **A:** Relationship between date of conception and date of sample collection for discovery (top) and replication (bottom) cohorts**. B**: Modelled seasonal change in methylation for 1,861 discovery CpGs in the discovery (top) and replication (bottom) cohorts. 60 ME CpGs are marked in red and the remaining 1,801 in blue. **C:** Conception date of modelled methylation maximum in each cohort for 125 replicating SoC-CpGs (left) and matched controls (right). **D:** Distribution of mean DNAm values (data from both cohorts combined) at i) SoC-CpGs that are putative MEs (n=10); ii) other SoC-CpGs (n=115); and iii) array background. **E:** Distribution of SoC-CpGs, matched controls and array background with respect to CpG islands (left) and gene locations (right). Error bars are bootstrapped 95% CIs. N / S Shore / Shelf: North / South Shore / Shelf respectively (regions proximal to CpG Islands defined in Illumina manifest).

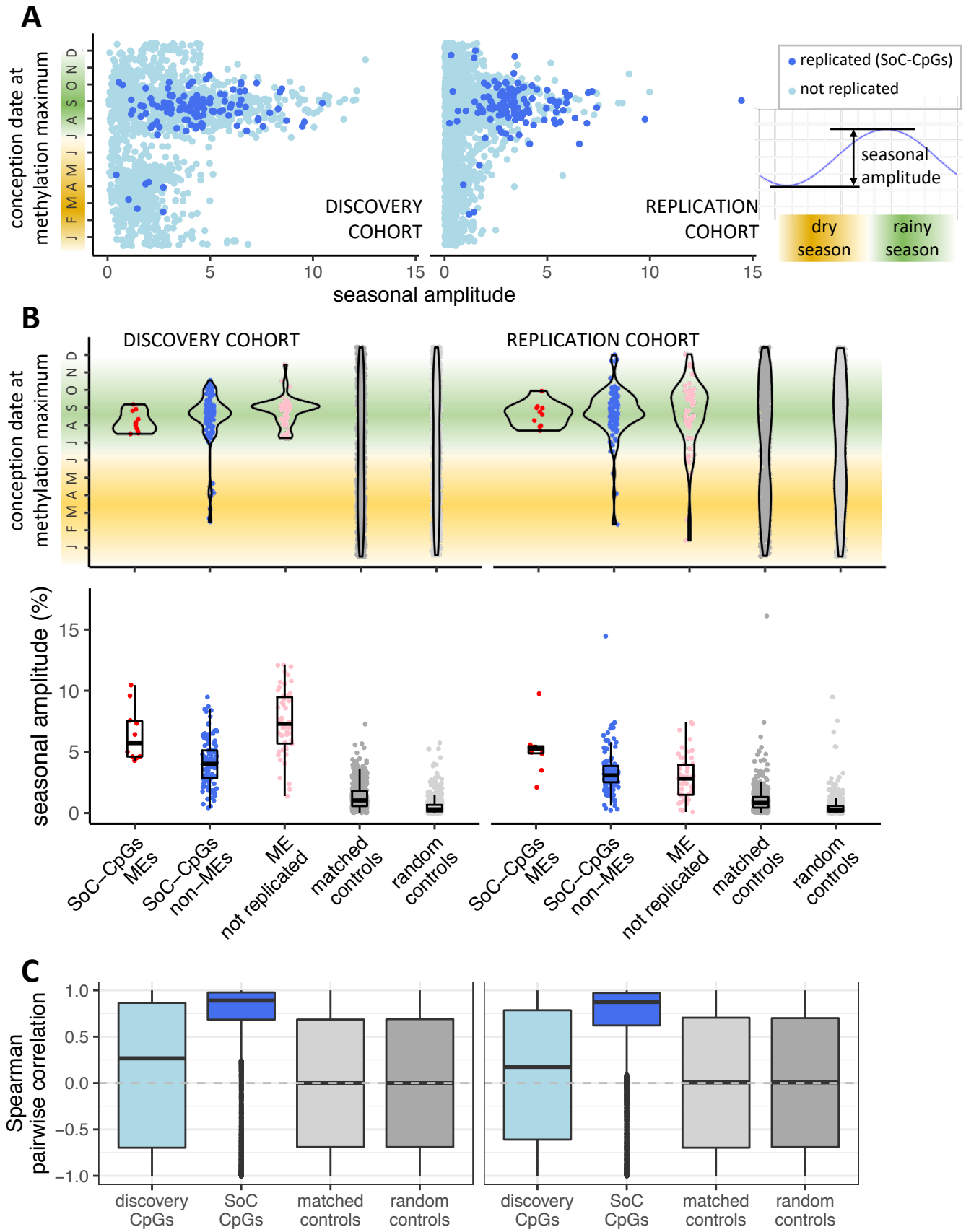
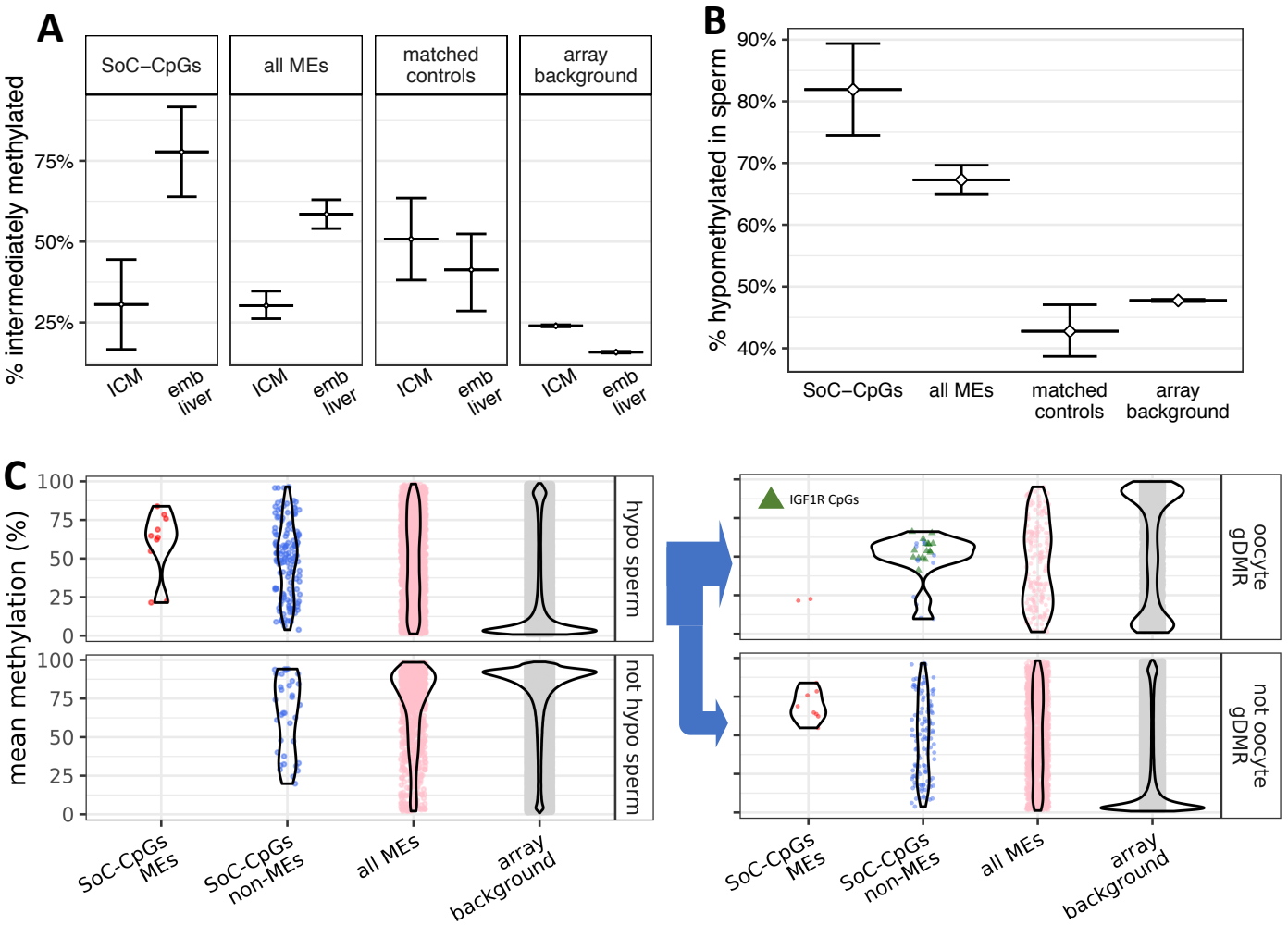


Figure 3. Association of periconceptional environment with DNA methylation 2. A: Date of modelled DNAm maximum vs seasonal amplitude in each cohort for (replicating) SoC-CpGs (n=125, dark blue) and non-replicating discovery CpGs (n=1,736, light blue). Seasonal amplitude is defined as the distance between modelled methylation peak and nadir (see inset). **B:** Date of conception at modelled methylation maxima (top) and seasonal amplitude (bottom) for i) SoC-CpGs that are known MEs (red); ii) other replicating CpGs (blue); iii) non-replicating MEs in discovery set (pink) and iv) matched and random controls (dark/light grey respectively). Green and yellow bands indicate the extent of the rainy and dry seasons respectively. **C:** Distribution of pairwise Spearman correlations for CpG sets in discovery (left) and replication (right) cohorts.



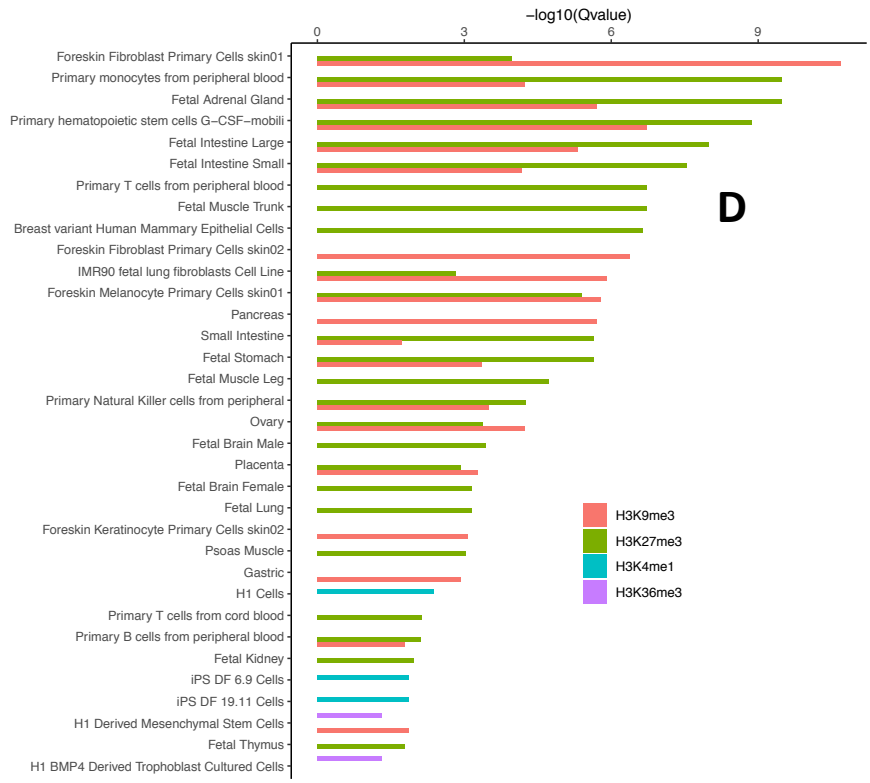


Figure 4. DNAm in sperm and the early embryo and overlap with histone marks. A: Proportion of intermediately methylated (10-90%) sites in pre-gastrulation inner cell mass (ICM) and post-gastrulation embryonic liver (emb liver) tissues, measured in RRBS embryo methylation data from Guo *et al.*[4]. Data comprises 67,870 CpGs covered at >=10x in both ICM and emb liver that overlap array background, including 36 SoC-CpGs, 470 ME and 63 matched control CpGs. Error bars represent bootstrapped 95% confidence intervals. **B:** Proportion of hypomethylated sites (methylation <10%) using sperm WGBS data from Okae *et al.*[21] Data comprises 294,240 CpGs covered at >10x and including 94 SoC-CpGs, 1,397 ME and 491 matched control CpGs. Bootstrapped CIs as above. **C:** (Left) Mean methylation at SoC-CpGs and controls, measured across all n=522 individuals in both cohorts, stratified according to whether loci are hypomethylated (top) or not hypomethylated (bottom) in sperm in the Okae *et al.* dataset. (Right) as left but further stratified according to whether loci are identified as oocyte gDMRs by Sanchez-Delgado *et al*[26] (top) or not (bottom). CpGs mapping to the *IGF1R* replicating region are marked as green triangles. **D:** Enrichment of histone (H3) marks assessed by eForge[28]. Bars represent $-\log_{10}$ enrichment FDR *q*-values for H3 marks derived from 39 tissues and cells analysed by the Roadmap Epigenomics Consortium. Only H3 marks, tissues and cells with significant enrichment (FDR $q \leq 0.05$) are shown.

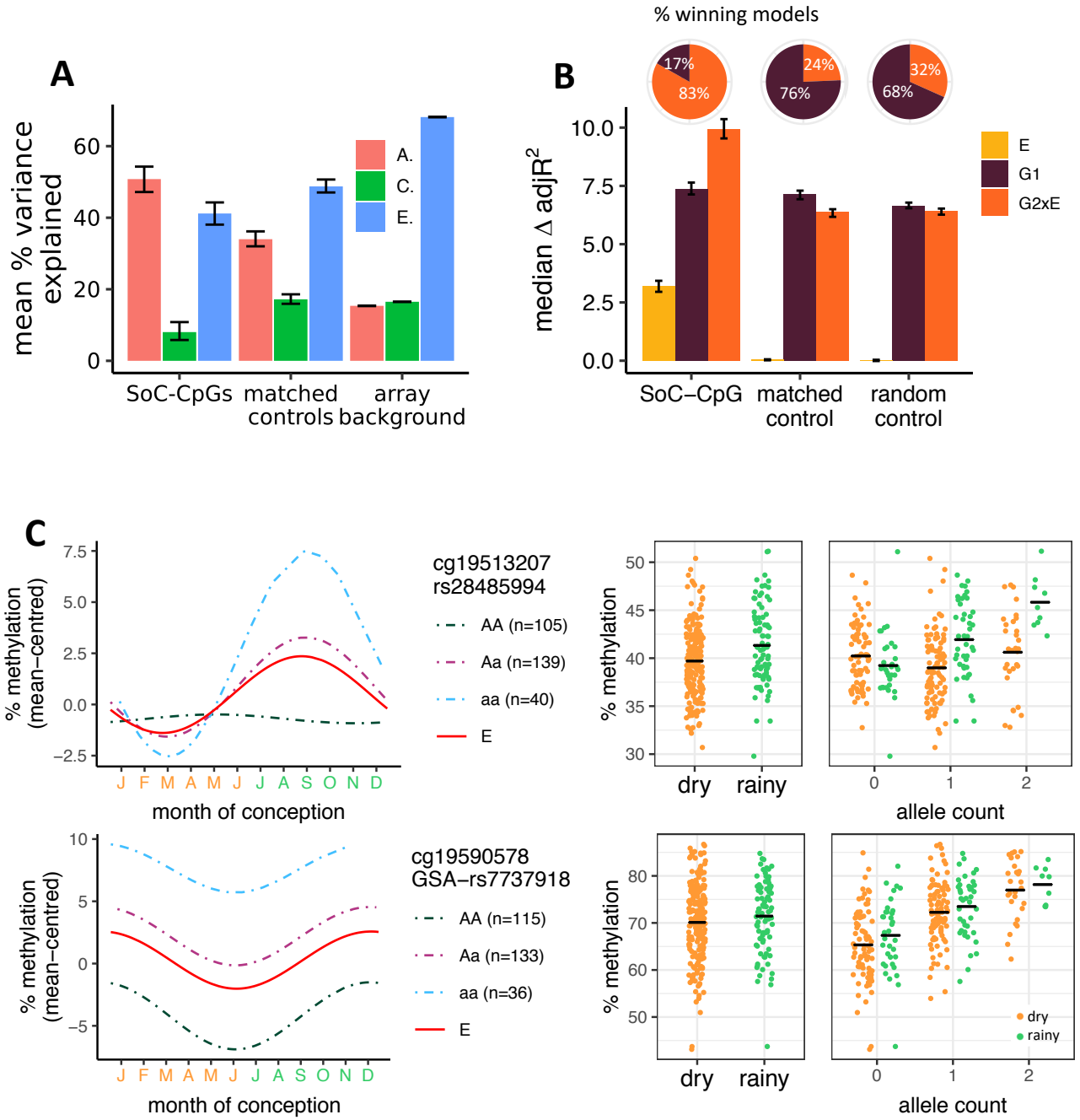


Figure 5. Influence of genotype, periconceptional environment and gene-environment interactions on DNAm. A: Methylation variance attributable to additive genetic (A), common (C) and non-shared (E) environment effects for replicated CpGs, matched controls and array background. Estimates for CpGs on the Illumina 450k are from Hannon et al[33]. Error bars represent bootstrapped 95% confidence intervals for the mean. **B:** Proportion of methylation variance explained by E, G1 and G2xE models for SoC-CpGs, matched and random control CpGs. Δ adjR^2 is the additional variance explained by the specified model, over and above a covariate-only model (see Methods). Pie charts show proportion of winning models, assessed using AIC. Note that E-only is never the winning model. **C:** (Left) Examples of SoC-CpGs with GxE (top) and G (bottom) winning models. Illumina CpG and rs identifiers for the most significant SNP are shown. Curves show Fourier regression model fitted values for E-only model (solid red line) for all individuals, and for individuals stratified by genotype (dashed lines). A/a major/minor alleles. (Right) Scatter plots of DNAm adjusted for baseline covariates only, stratified by season of conception (left) and additionally stratified by minor allele count (right). For ease of visualisation, seasons are dichotomised: dry season=Jan-Jun (orange); rainy season=Jul-Dec (green). Black horizontal lines are stratified mean values.