Environmentally sensitive hotspots in the methylome of the early human embryo

Silver MJ[1*†], Saffari A[1], Kessler NJ[2], Chandak GR[3], Fall CHD[4], Issarapu P[3], Dedaniya A[3], Betts M[1], Moore SE[1,5], James PT[1], Monk D[6,7] and Prentice AM[1*]

Author Affiliations

[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 authors
[†]Lead contact: 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 134 loci with a common season-of-conception methylation signature in two independent child cohorts. These robust candidates for sensitivity to early environment were highly enriched for metastable epialleles, parent-of-origin specific methylation and regions hypomethylated in sperm. 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 genomic remethylation in the very early embryo could thus constitute a sense-record-adapt mechanism linking early environment to later phenotype.

INTRODUCTION

DNA methylation (DNAm) plays an important role in a diverse range of epigenetic processes in mammals including X-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[2,3].

There is extensive remodelling of the human methylome in the very early embryo when parental gametic methylation marks are extensively erased before acquisition of tissues-specific marks at implantation, gastrulation and beyond[4]. Given these widespread changes in the early methylome, the days following conception may be a window of heightened sensitivity to the external environment, potentially stretching back to the period before conception coinciding with late maturation of oocytes and spermatozoa for methylation marks that (partially) evade periconceptional reprogramming[5].

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 exposure-related postnatal effects and reverse causation[6].

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[7,8]. We interrogate early embryonic events by focussing on metastable epialleles (MEs). First identified in isogenic mice, MEs exhibit stable patterns of systemic (cross-tissue) inter-individual variation (SIV) indicating stochastic establishment of methylation marks prior to gastrulation when tissue differentiation begins[9], and several MEs have been shown to be sensitive to periconceptional nutrition in mice[10]. Human MEs thus serve as a useful tool for studying the effects of the early environment on DNAm, by enabling the use of easily biopsiable tissues (e.g. blood) that can serve as a proxy for systemic methylation, and by pinpointing the window of exposure to the periconceptional period.

In this study we assess the influence of seasonality on DNAm in two Gambian child cohorts[11,12], enabling robust identification of loci showing consistent effects at the ages of 24 months and 8-9 years. 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 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, gametic methylation and the establishment of DNAm states in early embryonic development.

The developmental origins of health and disease (DOHaD) hypothesis posits the existence of mechanisms linking prenatal nutrition to lifelong metabolic disease[5]. It has also been proposed 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[13]. 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 and Figure 1A. 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 ENID ('discovery') and EMPHASIS ('replication') cohorts respectively. We modelled the effect of date of conception on DNAm using Fourier regression[14] which makes no prior assumptions about specific seasonal windows driving DNAm changes in offspring (see Methods).

We began by identifying 2,091 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 both cohorts (Fig. 1B, Supplementary Table 1). Next, we identified a subset of 134 'SoC-CpGs', defined as CpGs from the discovery CpG set with an FDR < 10% in the replication cohort (Supplementary Table 2).

SoC-CpGs showed a highly consistent seasonal pattern across both cohorts (Fig. 1C; Pearson correlation R=0.59, $p=7.7\times10^{-14}$ for conception date of modelled methylation maximum). 60% of SoC-CpGs exist as singletons, defined as having no SoC-CpG within 1,000bp, and 85% fall within clusters of 4 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 and high variance controls, SoC-CpGs are highly enriched for intermediate methylation states, most notably at 10 MEs previously identified in multi-tissue screens in adult Caucasians (Figure 1D; Supplementary Table 5; see Table 2 for details of ME and control loci considered). SoC-CpGs are enriched at CpG islands but depleted in open sea and 5' untranslated regions (proximal to transcriptions start sites) compared to controls (Fig. 1E).

In the discovery cohort, SoC-CpGs and non-replicating discovery CpGs show a distinct pattern of methylation maxima for conceptions falling within the July-September period (Fig. 2A left). This pattern is particularly marked at SoC-CpGs in both cohorts (Figs. 2A and 2B top), and also at MEs generally (even if non-replicating) (Fig. 2B top). The July-September period corresponds to the peak of the Gambian rainy season, a strong validation of previous Gambian studies in babies and infants that focussed on conceptions at peak seasons only, with similar observations of increased methylation in conceptions at the peak of the rainy season compared to peak dry season[15–17]. Methylation minima fall within the February-April period, corresponding to the peak 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. 2A and 2B bottom; Supplementary Table 6; Wilcoxon Rank-Sum test p-value ranging from 1.2×10^{-7} to 5.7×10^{-72}). Furthermore, there is evidence of a substantial and significant decrease in seasonal amplitude at non-replicating MEs in the older cohort (Fig. 2B bottom; median amplitude decrease=4.4%; Wilcoxon $p=6.8\times10^{-13}$), and a small, but significant decrease at SoC-CpGs that are not known MEs (median decrease=1.0%; $p=2.2\times10^{-5}$; Supplementary Table 6). There is no corresponding amplitude decrease in replicating MEs, or in either control set.

Compared to array background, both discovery and SoC-CpG sets are highly enriched for MEs (approximately 6-fold in discovery set and 17-fold in set of SoC-CpGs; $p=3.5x10^{-28}$ and $1.4x10^{-9}$ respectively), whereas no significant ME enrichment is observed at high variance CpGs (Supplementary Table 7). Finally, intra-individual methylation states are highly correlated at a large majority of SoC-CpGs in both cohorts, in marked contrast to discovery CpGs and controls (Fig. 2C). As expected, pairwise correlations at the minority of SoC-CpGs showing a strong negative intra-individual correlation consist largely of a small number of loci with methylation maxima in dry season conceptions (as shown in Fig. 1C).

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 MEs identified in a whole-genome bisulfite-seq (WGBS) multi-tissue screen[18]. We aligned our data with public reduced representation bisulfite-seq (RRBS) data from human IVF embryos[4] 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 at SoC-CpGs in post-gastrulation embryonic liver tissue. This strongly contrasted with a general trend of genome-wide hyper- and hypo-methylation at highly variable CpGs and at loci mapping to array background (Fig. 3A). We observed a similar pattern at MEs (Fig. 3A; all 1,881 ME CpGs irrespective of their association with SoC – see Table 2).

We previously observed consistent hypomethylation at ME loci across all gametic and early embryonic developmental stages, most notably in sperm[18]. We tested the latter observation at SoC-CpGs by aligning our data with public sperm WGBS data[19], restricting our analysis to the 389,360 CpGs mapping to array background that were covered at ≥ 10x. All 134 SoC-CpGs were covered in the WGBS dataset and these showed a marked decrease in sperm methylation, with 83% [76-90%] of replicated loci hypomethylated (methylation <10%) in sperm, compared to 49% [48%-50%] and 48% [48-48%] at loci mapping to highly variable CpGs and array background respectively (Fig. 3B; brackets are bootstrapped 95% CIs). This strong enrichment for sperm hypomethylation was also observed at ME CpGs (Fig. 3B). Interestingly, postnatal intermediate methylation states at SoC-CpGs were preserved in both Gambian cohorts irrespective of putative sperm methylation states, in contrast to loci mapping to high variance and array background CpGs where methylation distributions strongly reflected sperm hypomethylation status (Fig. 3C 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[20–22], 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[23]. 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 (41- and 15-fold enrichment, $p=4.2x10^{-12}$ and $1.8x10^{-36}$ respectively; Fig. 3D, green bars; Supplementary Table 8). PofOm enrichment at SoC-CpGs is partially driven by a large (8 CpG) region at *IGF1R*, and a single replicating ME-CpG proximal to the human imprinted 14q32 region (Supplementary Table 2). We also found evidence of significant but smaller PofOm enrichment amongst high variance CpGs (Fig. 3D; Supplementary Table 8).

Regions of PofOm detected in postnatal samples tend to be differentially methylated in gametes[23], and may thus have evaded the widespread epigenetic reprogramming that occurs

in the pre-implantation embryo[24]. 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[25]. 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, taking methylation variability into account. We began by observing a very large enrichment for oocyte (maternally methylated) gDMRs, but not sperm gDMRs, at all PofOm loci identified by Zink *et al.*[23] (Fig. 3D, right; Supplementary Table 8), confirming previous observations of an excess of PofOm loci that are methylated in oocytes only[23]. We found a particularly strong 122-fold enrichment for oocyte gDMRs (oo-gDMRs) persisting in placenta. We next interrogated SoC-CpGs and MEs and again found evidence for strong enrichment of oocyte, but not sperm gDMRs, at these loci (5.6-fold enrichment for oo-gDMRs, $p=6.4x10^{-8}$; Fig. 3D, left; Supplementary Table 8). Once again this enrichment was particularly marked at oo-gDMRs persisting in placenta (11-fold enrichment, $p=2.5x10^{-8}$; Fig. 3D, left; Supplementary Table 8). As with sperm hypomethylation, enrichment of oo-gDMRs at SoC-CpGs is partially driven by multiple CpGs at *IGF1R* (Supplementary Table 2). The 14q32 ME-CpG is not classified as an oo-gDMR by Sanchez-Delgado *et al*.[25] (Fig. 3C bottom right), although it shows approximately 50% methylation in both Gambian cohorts as expected for loci with PofOm. 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. 3C, bottom right), suggesting that regional sperm hypomethylation is a key factor associated with sensitivity to periconceptional environment at these loci.

Enrichment of transposable elements and transcription factors associated with genomic imprinting

Variable methylation states at MEs are associated with neighbouring transposable elements (TEs) in murine models[26,27], and we have previously observed enrichment for specific proximal endogenous retroviruses (ERV1 and ERVK) in screens for human MEs[18,22]. Here we found evidence for enrichment of human ERV1 and ERVK at SoC-CpGs, and for 4 ERV classes at MEs mapping to array background: ERV1, ERVK, ERVL and ERV-MaLR (Supplementary Table 7). We also observed significant enrichment for these ERV classes amongst highly variable loci (Supplementary Table 7).

Enrichment for PofOm 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) linked to such mechanisms[28]. Here we found no evidence for enrichment of these TFs at SoC-CpGs, but we did find evidence of enrichment for proximal CTCF and ZFP57 binding at MEs mapping to array background (Supplementary Table 7), partially confirming ME enrichment for these TFs at ME loci on Illumina arrays.

Influence of genotype and gene-environment interactions

Genetic variation, primarily in *cis*, is a major driver of inter-individual variation in DNAm[29], and there is evidence that CpG sites with systemic inter-individual variation (SIV), a hallmark of MEs, show higher levels of heritability[30].

A previous analysis quantified methylation variance explained by additive genetic variation, and common and non-shared environment in 1,464 twin pairs from the British E-Risk study[30]. We began by reproducing the result from Hannon *et al*.[30] that non-shared environment (which includes measurement error) explains the major part of methylation variance in array background (Fig. 4A). We further found a marked increase in methylation variance explained by additive genetic effects at SoC-CpGs and high variance CpGs on the array (Fig. 4A). On the assumption that the former will be enriched for MEs exhibiting 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[30].

We next directly explored the influence of genotype and environment at SoC-CpGs in the EMPHASIS (replication) cohort, for which we had data on 293 individuals measured at 286,552 polymorphic variants on the Illumina Global Screening Array (GSA)[12]. Following a similar strategy to that used in a recent study in the GUSTO cohort[31], we performed a screen for genetic effects on DNAm. We began by analysing genome-wide SNP-DNAm associations at SoC-CpGs, and at a random sample of array background and high-variance control CpGs (see Table 2) to identify putative methylation quantitative trait loci (mQTL). We also tested for gene-environment interactions (GxE) on DNAm (see Methods for further details).

We next selected the most significant ('winning') mQTL (G1) and GxE (G2) SNP for each SoC-CpG (Supplementary Table 9). 8.2% of SoC-CpGs had a significant associated mQTL (FDR<10%). No significant GxE associations were identified, although it should be noted that this analysis had reduced power to detect these. A small number of winning G1 and G2 SNPs mapped to two CpGs (G1: 6 SNPs; G2: 3 SNPs), but none mapped to more than two (Supplementary Table 10). Furthermore, no pattern of SNP clustering was discernible (Supplementary Fig. 4). Together these observations suggest that DNAm at SoC-CpGs is not primarily driven by genetic variants at any specific locus covered by the GSA.

To assess the potential for genetic confounding of SoC-associated DNAm patterns, we tested each winning G1 and G2 SNP for association with SoC using 5 different genotypic models. Accounting for the number of SNPs tested, only one of the 258 unique G1 and G2 SNPs passed a Bonferroni-adjusted significance threshold (G1 SNP: rs11922293; Supplementary Table 11). This mQTL SNP was associated with a single SoC-CpG and was not close (<10Mbp) to any other winning G1 SNP, confirming no evidence of widespread confounding of SoC-associated DNAm by genetic loci on the GSA.

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) models (see Methods for further details). Results for SoC-CpGs were compared to randomly selected high variance and control CpGs (see Table 2). 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. 4B, Supplementary Table 12). 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 94% of SoC-CpGs, compared with 33% and 29% for random and high variance controls respectively (Fig.4B inset).

As expected, year-round DNAm at a CpG where G1 is the winning model indicates a strong mQTL effect on mean methylation (Fig. 4C, bottom left). In contrast, at a CpG where GxE effects dominate, the strength of the seasonality effect is modified by genotype (Fig. 4C, top left; dashed lines); revealing a strong seasonal pattern that is not apparent when modelling data unstratified by genotype (same figure, solid red line). Scatter plots of underlying individual-level DNAm data adjusted for baseline covariates support these observations (Fig. 4C 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[31]. 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*.[31] that overlapped array background. We observed a highly significant 17-fold enrichment of these GxE CpGs amongst SoC-CpGs ($p=1.6x10^{-05}$; Supplementary Table 7). We found a smaller, but still highly significant 7-fold enrichment amongst the much larger set of the top 5% of CpGs by methylation variance ($p=1.9x10^{-97}$), suggesting that enrichment for Teh *et al.* GxE CpGs amongst SoC-CpGs is not purely driven by Teh *et al.*'s focus on highly variable CpGs.

DISCUSSION

We have exploited a natural 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 134 '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, suggest there may be many more hotspots sensitive to periconceptional environment across the human methylome.

This analysis contrasts with 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[15–17,22,32].

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 human MEs and related loci with characteristic methylation signatures suggestive of establishment early in embryonic development[17,18]. Second, like MEs, season-associated loci exhibit highly unusual methylation dynamics in early stage embryos[8]. Third, also in common with MEs, they have distinctive gametic methylation patterns, notably an increased proportion displaying hypomethylation in sperm[18].

Greater methylation in offspring conceived at the peak of the Gambian rainy season is consistent with previous findings at putative MEs and correlated regions of SIV in this population[15–17,22,32,33]. This observation is now greatly strengthened by the application of Fourier regression to model the effect of periconceptional environment in year-round conceptions – an approach that makes no prior assumption of where methylation peaks and nadirs may occur. The number of identified SoC-CpGs is also substantially increased in this

study, and comparisons with high variance and array-wide control CpGs increase confidence that these findings are not statistical artefacts.

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. In particular evidence of an attenuation of SoC effects at known MEs in older children suggests that screens for MEs in adults used in this analysis are likely to have missed signatures of metastability that are present in early postnatal tissues. SoC effect attenuation could also explain the lack of replication of SoC associations at the majority of CpGs from the discovery set. 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[34,35].

Methylation states at SoC-associated loci which are distributed throughout the genome are highly correlated within individuals, 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[36]), where no intra-individual correlation between stochastic methylation states was observed, although it is important to note that the mice were not exposed to different environments[27].

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 parent-of-origin-specific methylation 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[20,21]. 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*[22], consistent with evidence of 'polymorphic imprinting' linked to prenatal environment at this locus[23,37]. 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[18], although we were unable to replicate this at SoC-associated loci identified in this study. This might 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.

Second, a feature of SoC-CpGs, including those with no evidence of PofOm, is a strong enrichment for ERV elements, most notably ERV1 and ERVK. This was also observed at MEs on the Illumina array, confirming our previous observations[18,22]. Enrichment of ERVs at SoC-CpGs is notable since most environment-sensitive mouse MEs are associated with IAPs (which are rodent-specific ERVs)[27], 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[38]. 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[20,39]. We previously identified a putative SoC-associated DMR in the *ZFP57* promoter in blood from Gambian infants[22], 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[17]. 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-associated loci is highly enriched for intermediate methylation states, in strong contrast to array-wide CpG methylation and to high variance CpGs in the discovery cohort. Intermediate methylation at MEs is observed in Gambians and in non-Africans[15,16,40–42], and this coincides with a similar observation at MEs in post-gastrulation embryonic tissues[18]. This latter observation includes measurements from single conceptuses, with methyl-seq read-level analyses indicating that intermediate methylation is driven by extended regions of variegated methylation states within an individual[18].

Taken together, the above evidence suggests that a periconceptional environmental exposure may perturb methylation by nudging the ratio of methylated to unmethylated alleles at hotspots of variegated and/or parent-of-origin-specific methylation 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.*[23] 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[43,44]. Interestingly, 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[45]. IGF1R signalling is implicated in fetal growth, glucose metabolism and cancer[46–48], and DNAm differences at *IGF1R* have been observed in birthweight-discordant adult twins[49]. Another SoC-associated locus with PofOm is approximately 300kbp from the 14q32 *DLK1-MEG3* imprinted region, close to the imprinted C14MC microRNA cluster[50], and within 80kb of a region with PofO-ASE[23]. Epigenetic and transcriptional changes at several C14MC microRNAs have been implicated in cancer[51–53], and genetic and epigenetic mutations in the 14q32 region are linked to imprinting disorders including Temple syndrome[54,55].

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[42]. Hypermethylation at the VMR reduces *POMC* expression by interfering with P300 TF binding at the intron2/exon3 boundary of the gene[56], and is linked to the presence of a primate-specific *Alu* element (transposon)[57]. This region has previously been associated with SoC and certain C1 metabolites in Gambian infants[42], and is associated with obesity in children and adults[42,56]. 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 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[58]. One could then speculate that SoC-CpGs, as loci with evidence of SoC effects that persist into later childhood (i.e. that replicate in the older cohort), are specifically designed for that purpose.

DNAm is strongly 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 genetic associations driving inter-individual methylation differences at multiple SoC-associated loci in *cis* or *trans*.

We did however uncover interesting 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 was constrained by reduced power and by limited coverage of the genotyping array, confidence in this observation is increased through our comparison of genetic effects at SoC-CpGs with high variance and array-wide controls. The potential dominance of GxE effects 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[31]. Widespread GxE interaction effects could manifest through the action of environmental factors on gene variant-associated transcription factors, although once again we found no evidence of clustered genetic variants driving these effects at multiple SoC-CpGs.

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[18], 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 high-resolution genotyping 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[59,60], and in genetic variation as a potential driver of methylation variance[61]. 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.

A further 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[13]. 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[5].

Materials and Methods

Gambian cohorts and sample processing

Detailed descriptions of the Gambian cohorts analysed in the season of conception study are published elsewhere[11,12]. 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[11]. DNA was extracted, bisulfite-converted and hybridised to Illumina HumanMethylation450 (hereafter "HM450") arrays following standard protocols (see Van Baak *et al.*[17] for further details). For the older cohort, DNA was extracted from blood samples from 289 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[12], and was bisulfite-converted and hybridised to Illumina Infinium Methylation EPIC (hereafter "EPIC") arrays, again using standard protocols.

Methylation array pre-processing and normalisation

Raw intensity IDAT files from the HM450 and EPIC arrays were processed using the *meffil*[62] package in R 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, correlation between technical replicates); 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[63], followed by Functional Normalisation to reduce technical variation based on principal component analysis of control probes on the arrays[64]. 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[14,65]. This models the relationship between a response variable (here DNAm) and a cyclical predictor (here date of conception). The latter is considered cyclical, since 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 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:

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

Where, for each individual i and CpG j:
M_{ij} is the logit-transformed methylation Beta value[66];
α_{ik} is the k_{th} 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 (i.e. n=1), this gives a sinusoidal pattern of variation, with a single maximum and minimum whose phase (position) 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 both cohorts, adjustment 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 latter was used to account for unmeasured and measured technical variation (due to bisulfite conversion sample plate, array slide etc) and for cell composition effects (see Supplementary Tables 13 and 14). 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 13). Child age was included as an additional adjustment covariate for the EMPHASIS cohort, since child ages ranged from 8 to 9 years, plus maternal nutritional intervention group (see Chandak *et al.*[12] for further details).

For CpG *j*, coefficient estimates β_j, γ_j and p-value p_j were determined as follows:

1. Fit model with a single pair of Fourier terms (n=1) using *lm()* in R to obtain estimates for β_1 and γ_1. Determine model goodness-of-fit by likelihood ratio test (LRT using *lrtest*() in R) by comparing the full model with a baseline model containing adjustment covariates only. Determine model p-value (p_1) from the corresponding LRT chi-squared statistic.
2. If $p_1 < 0.05$, fit a second model with two pairs of Fourier terms (n=2) to obtain estimates β_{1*}, β_{2*} and γ_{1*}, γ_{2*}. Determine model goodness-of-fit by LRT comparing this model with the model in 1. above. Model p-value (p_2) is the corresponding LRT chi-squared statistic.
3. If $p_1 \geq 0.05$: $p_j = p_1$; and select model with coefficients $\beta_j = \beta_1$, $\gamma_j = \gamma_1$
   if $p_1 < 0.05$:
   if $p_2 < 0.001$: $p_j = p_2$; and select model with coefficients $\beta_j = \beta_{1*}$, β_{2*} and $\gamma_j = \gamma_{1*}$, γ_{2*}
   if $p_2 \geq 0.001$: $p_j = p_1$; and select model with coefficients $\beta_j = \beta_1$, $\gamma_j = \gamma_1$

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

From the analysis in the discovery (ENID) cohort, CpG p-values as described above 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 the Benjamini & Hochberg method (*p.adust()* in R with *method*='fdr'). 2,091 CpGs had a FDR<10% and 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. After accounting for 2,091 multiple tests, 134 CpGs had a FDR<10% and these formed the set of replicating 'SoC-CpGs'.

Additional modelling of seasonal variation in blood cell composition

Seasonal variation in blood cell composition was modelled by Fourier regression with sex (ENID+EMPHASIS) and age (EMPHASIS only) as adjustment covariates. Cell count estimates using the Houseman method[67] were obtained using the *estimateCellCounts*() from *minfi* in R. In each case the best fitting model with one or two pairs of Fourier terms was determined by LRT. Best fit models indicated no consistent or marked seasonal differences within and between cohorts (Supplementary Figure 5).

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, all of which have evidence of systemic inter-individual variation of DNAm with establishment in the early embryo: putative MEs identified in a multi-tissue WGBS screen in Kessler et al.[18]; and CpGs exhibiting 'epigenetic supersimilarity' and/or SIV described in Van Baak *et al.*[17].
ii. *699 parent-of-origin-specific CpGs (PofOm CpGs)* overlapping 229 regions with PofOm identified in Supplementary Table 1 from Zink *et al.*[23].
iii. *889 GxE CpGs* listed in Teh *et al.*[31] Supplementary Table 6. These are highly variable loci where methylation variance is best explained by GxE models, with E covering a range of *in utero* exposures.

Early stage embryo and sperm methylation data

RRBS methylation data from Guo *et al.*[4] 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.*[18].

Sperm methylation data from Okae *et al*.[19] was downloaded from the Japanese Genotype-phenotype Archive (accession number S00000000006); only CpGs covered at ≥ 10x were considered in this analysis.

Germline gDMRs

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

Transposable elements and transcription factors

Transposable element regions determined by RepeatMasker were downloaded from the UCSC hg19 annotations repository. Further details on these and ZFP57, TRIM28 and CTCF transcription factor binding sites in human embryonic kidney and human embryonic stem cells used in this analysis are described in Kessler et al.[18].

Gene and Gene-environment interaction analyses

Gene-DNAm association analysis were performed on all 289 individuals from the EMPHASIS (replication) cohort, since this was the only cohort with associated genetic data. 134 SoC-CpGs, plus a random sample of 2,091 array background and high variance controls were considered in this analysis (see Table 2). Genotypes were obtained from the Illumina Infinium Global Screening Array-24 v1.0 Beadchip (Illumina, California, U.S.) following standard protocols[68], with 642,824 SNPs available for analysis after QC of which 286,552 were polymorphic in this dataset. 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 174,489 SNPs with a minimum minor allele frequency (MAF) of 11.4%.

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

Environment (E), and genome-wide genetic (G) and gene-environment (GxE) associations were assessed using the *GEM* package from Bioconductor[69], following a similar strategy to that described in Teh *et al.*[31].

In total, 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 (sinθ, cosθ) and covs corresponding to the same adjustment covariates used in the main analysis.

2. G-model:

   M_j ~ covs + cosθ + G

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

   OR

   M_j ~ covs + sinθ + G

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

   Here, G is SNP genotype coded as allelic dosage (0,1,2) and covs are adjustment covariates as described above. In each case the less significant Fourier term from the E-model is included as an additional covariate to ensure unbiased comparison between E, G and GxE models.

3. GxE-model

   M_j ~ covs + cosθ + G + G x sinθ

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

   OR

   M_j ~ covs + sinθ + G + G x cosθ

when $\cos\theta$ is the most significant Fourier term in E-model.

Here, G and covs are as described above. Again, in each case the less significant Fourier term from the E-model is included as an additional covariate to ensure unbiased comparison between E, G and GxE models.

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 regeression models), comparisons of methylation variance explained by E, G1 and G2xE models (Figure 4b bar plots, Supplementary Table 12) are based on adjusted R-squared values. In each case, for each CpG

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

Where $\text{adjR}^2_{\text{model}}$ is the adjusted R^2 value for the full model, and $\text{adjR}^2_{\text{cov}}$ is the adjusted R^2 for the covariate-only model, including the less-significant Fourier term as described above. In the case of E-only models, the full model includes the most significant Fourier term (sin or cosine), and the covariate-only model includes all other model covariates including the less significant Fourier term.

Winning models (Figure 4b pie charts) are those with the lowest value of the Akaike Information Criterion (AIC[70]).

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

Acknowledgements

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), and we acknowledge the work of Z. Herceg, M. N. Routledge, Y. Y. Gong, and H. Hernandez-Vargas in acquiring this data. 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). We acknowledge the work of the full EMPHASIS Study Group (www.emphasisstudy.org) in acquiring this data.

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. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610 (2014).
5. Fleming, T. P. *et al.* Origins of lifetime health around the time of conception: causes and consequences. *Lancet* **391**, 1842–1852 (2018).
6. Birney, E., Smith, G. D. & Greally, J. M. Epigenome-wide Association Studies and the Interpretation of Disease -Omics. *PLOS Genet.* **12**, e1006105 (2016).
7. 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).
8. 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).
9. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* **18**, 348–51 (2002).
10. 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).
11. 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).
12. 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).
13. 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).
14. 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).
15. Waterland, R. A. *et al.* Season of conception in rural gambia affects DNA methylation at putative human metastable epialleles. *PLoS Genet.* **6**, e1001252 (2010).
16. Dominguez-Salas, P. *et al.* Maternal nutrition at conception modulates DNA methylation of human metastable epialleles. *Nat. Commun.* **5**, 1–7 (2014).
17. Van Baak, T. E. *et al.* Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biol.* **19**, 2 (2018).
18. 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).
19. Okae, H. *et al.* Genome-Wide Analysis of DNA Methylation Dynamics during Early Human Development. *PLoS Genet.* **10**, e1004868 (2014).
20. 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.* (2019). doi:10.1038/s41576-018-0092-0
21. 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
22. 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).
23. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and transcriptomes. *Nat. Genet.* **50**, 1542–1552 (2018).
24. Meyenn, F. Von & Reik, W. Forget the Parents : Epigenetic Reprogramming in Human Germ Cells. *Cell* 1248–1251 (2015).
25. Sanchez-Delgado, M. *et al.* Human Oocyte-Derived Methylation Differences Persist in the Placenta Revealing Widespread Transient Imprinting. *PLOS Genet.* **12**, e1006427 (2016).
26. Waterland, R. A. & Jirtle, R. L. Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.* **23**, 5293–300 (2003).
27. 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
28. Kessler, N. J., Waterland, R. A., Prentice, A. M. & Silver, M. J. SUPP INFO Establishment of environmentally-sensitive DNA methylation states in the very early human embryo. *Sci. Adv. (in Press.* (2018). doi:10.1126/sciadv.aat2624
29. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* **17**, 61 (2016).
30. Hannon, E. *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.* **14**, 1–27 (2018).
31. 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
32. 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).
33. Gunasekara, C. J. *et al.* A genomic atlas of systemic interindividual epigenetic variation in humans. *Genome Biol.* **20**, 105 (2019).
34. 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).
35. Simpkin, A. J. *et al.* Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum. Mol. Genet.* **24**, 3752–3763 (2015).
36. 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).
37. 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
38. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease. *Nature* **571**, 489–499 (2019).
39. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the

evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
40. 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).
41. 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).
42. 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).
43. 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).
44. 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).
45. 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).
46. Randhawa, R. & Cohen, P. The role of the insulin-like growth factor system in prenatal growth. *Mol. Genet. Metab.* **86**, 84–90 (2005).
47. 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).
48. Larsson, O., Girnita, A. & Girnita, L. Role of insulin-like growth factor I receptor signalling in cancer. *Br. J. Cancer* **92**, 2097–2101 (2005).
49. 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).
50. 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).
51. 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).
52. 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).
53. 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).
54. 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).
55. 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).
56. Kuehnen, P. *et al.* An Alu Element–Associated Hypermethylation Variant of the POMC Gene Is Associated with Childhood Obesity. *PLoS Genet.* **8**, e1002543 (2012).
57. Kuehnen, P. & Krude, H. Alu elements and human common diseases like obesity. *Mob. Genet. Elements* **2**, 197–201 (2012).
58. Low, F. M., Gluckman, P. D. & Hanson, M. A. Developmental Plasticity, Epigenetics and Human Health. *Evol. Biol.* **39**, 650–665 (2012).

59. Webster, A. P. *et al.* Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. *Genome Med.* **10**, 1–12 (2018).
60. Tobi, E. W. *et al.* DNA methylation as a mediator of the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Sci. Adv.* **4**, eaao4364 (2018).
61. Ek, W. E. *et al.* Genetic variants influencing phenotypic variance heterogeneity. *Hum. Mol. Genet.* **27**, 799–810 (2018).
62. 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).
63. 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).
64. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15**, 0–42 (2014).
65. 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).
66. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11**, 587 (2010).
67. Jaffe, A. E. & Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
68. Guo, Y. *et al.* Illumina human exome genotyping array clustering and quality control. *Nat. Protoc.* **9**, 2643–2662 (2014).
69. 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).
70. Akaike, H. A New Look at the Statistical Model Identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974).

TABLES

Table 1. Gambian seasonality-methylation analysis: cohort characteristics

cohort	sample size	Age at sample collection	% 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	2,091	SoC-associated loci identified in the discovery cohort (FDR<10%)
SoC-CpGs	134	'Discovery CpGs' with significant seasonal variation in the replication cohort (FDR<10%)
Metastable epialleles (MEs)	1,881	ME CpGs identified in multi-tissue screens by Van Baak *et al*[17] and Kessler *et al*[18] overlapping array background
Parent-of-origin specific methylation	699	Parent-of-origin specific methylation loci identified in Zink *et al*[23] overlapping array background
GxE CpGs	889	CpGs with evidence of Gx(in utero)E interactions identified in Teh et al.[31] overlapping array background
High variance top 5%	18,281	CpGs in the top 5% by methylation variance in the discovery cohort
High variance Controls*	2,091	Random sample of 2,091 CpGs from the 'high variance top 5%' set
Random controls*	2,091	Random sample of 2,091 CpGs from array background

* random subsamples of equal size to the discovery CpG set. Used for computational tractability in modelling analyses

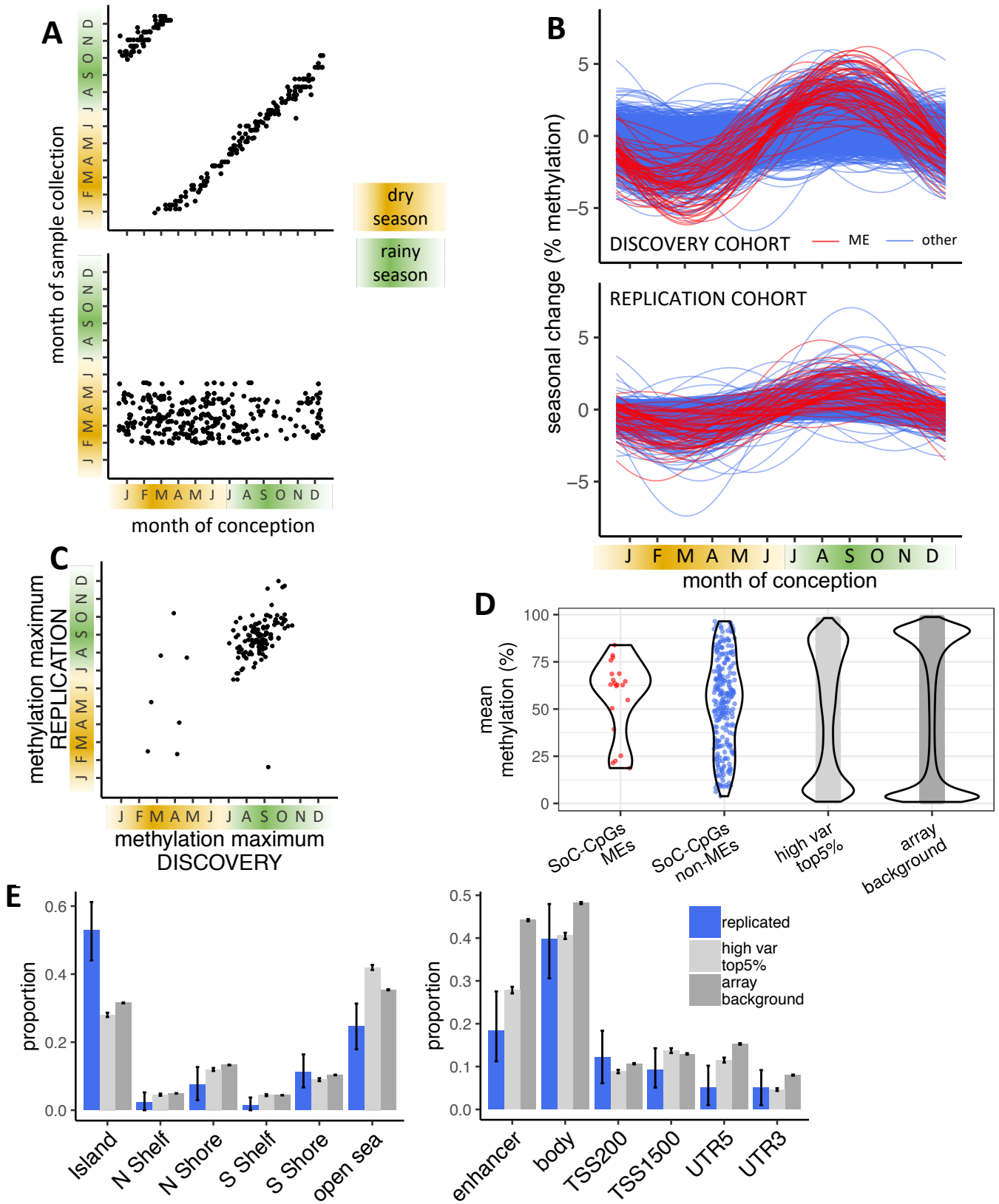

Figure 1. Association of periconceptional environment with DNA methylation 1. A: Relationship between date of conception and date of sample collection for discovery and replication cohorts in the Gambian seasonality-DNAm analysis**.** DNAm differences associated with season of conception are potentially confounded by season of sample collection in the discovery cohort (ENID - top), since samples are collected at age 2yrs. This is not the case in the replication cohort (EMPHASIS - bottom) where all samples are collected in the Gambian dry season. **B**: Modelled seasonal change in methylation for 2,091 discovery CpGs in the discovery (top) and replication (bottom) cohorts. 61 ME CpGs are marked in red and the remaining 2,030 in blue. **C:** Conception date of modelled methylation maximum in each cohort for 134 replicated CpGs. **D:** Distribution of mean DNAm values (data from both cohorts combined) at i) SoC-CpGs that are known MEs (n=10 marked in red); ii) other SoC-CpGs (n=124, blue); and iii) high variance and array background CpGs as controls (n=18,281 & 391,484; light/dark grey respectively). **E:** Distribution of SoC-CpGs and controls with respect to CpG islands (left) and gene locations (right). Error bars are bootstraped 95% CIs. N / S Shore / Shelf: North / South Shore / Shelf respectively (regions proximal to CpG Islands defined in Illumina manifest).

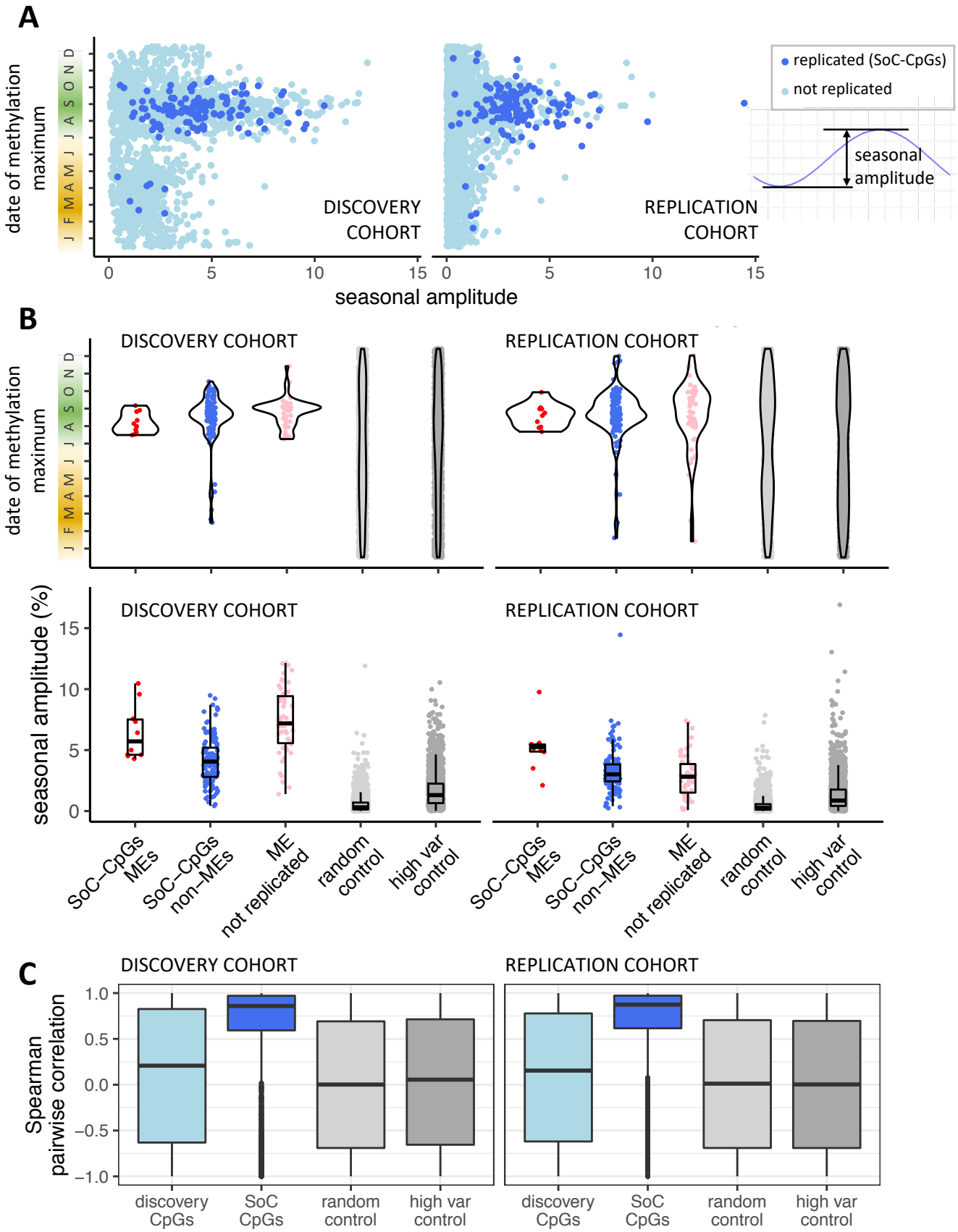

Figure 2. Association of periconceptional environment with DNA methylation 2. A: Date of modelled DNAm maximum vs seasonal amplitude in each cohort for (replicating) SoC-CpGs (n=134, dark blue) and non-replicating discovery CpGs (n=1,957, light blue). Seasonal amplitude is defined as the distance between 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) random and high variance controls (light/dark grey respectively). **C:** Distribution of intra-individual pairwise methylation correlations for CpG sets in discovery (left) and replication (right) cohorts.

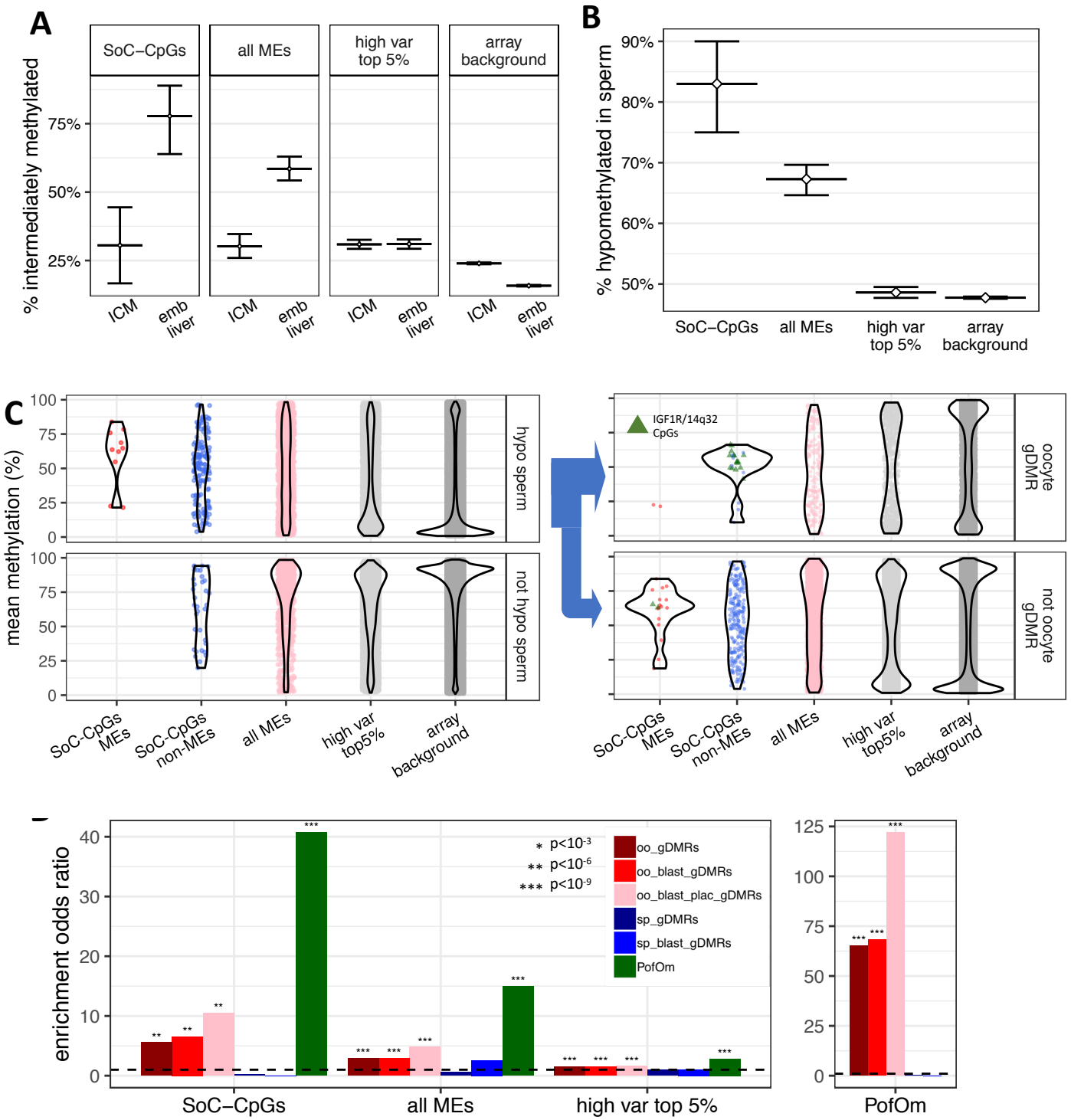

Figure 3. Early embryo DNA methylation dynamics at SoC loci. A: Proportion of intermediately methylated (10-90%) sites in pre-gastrulation inner cell mass (ICM) and post-gastrulation embryonic liver (emb liver) tissues, measured in RRBS embryo methylation data from Guo *et al.*[4]. Data comprises 67,870 CpGs covered at >=10x in both ICM and emb liver by Guo *et al.* that overlap array background, including 36 SoC-CpGs and 470 ME CpGs. Error bars represent bootstrapped 95% confidence intervals. **B:** Proportion of hypomethylated sites (methylation <10%) using sperm WGBS data from Okae *et al.*[19] Data comprises 389,360 CpGs covered at >10x and includes all 134 SoC-CpGs and 1,881 ME CpGs. Bootstrapped CIs as above. **C:** (Left) Mean methylation at SoC-CpGs and controls, measured across all n=522 individuals in both cohorts, sub-divided according to whether loci are hypomethylated (top) or not hypomethylated (bottom) in sperm in the Okae *et al.* dataset. (Right) as left but further sub-divided according to whether loci or oocyte gDMRs (top) or not (bottom). CpGs mapping to the *IGF1R* and 14q32 replicating regions are marked as green triangles. Note that the 14q32 CpG is also an ME. **D:** (Left) Enrichment of oocyte (oo=maternally methylated) and sperm (sp=paternally methylated) gDMRs[25] and PofOm[23] at SoC-CpGs, all MEs and CpGs in the top 5% by variance set. Analyses includes loci with PofOm that persists at the blastocyst (blast) and placental (plac) stages. (Right) positive control analysis demonstrating that PofOm loci identified in postnatal samples by Zink *et al.*[23] are very highly enriched for maternal gDMRs. Dashed horizontal is OR=1. Note different y-axis scales..

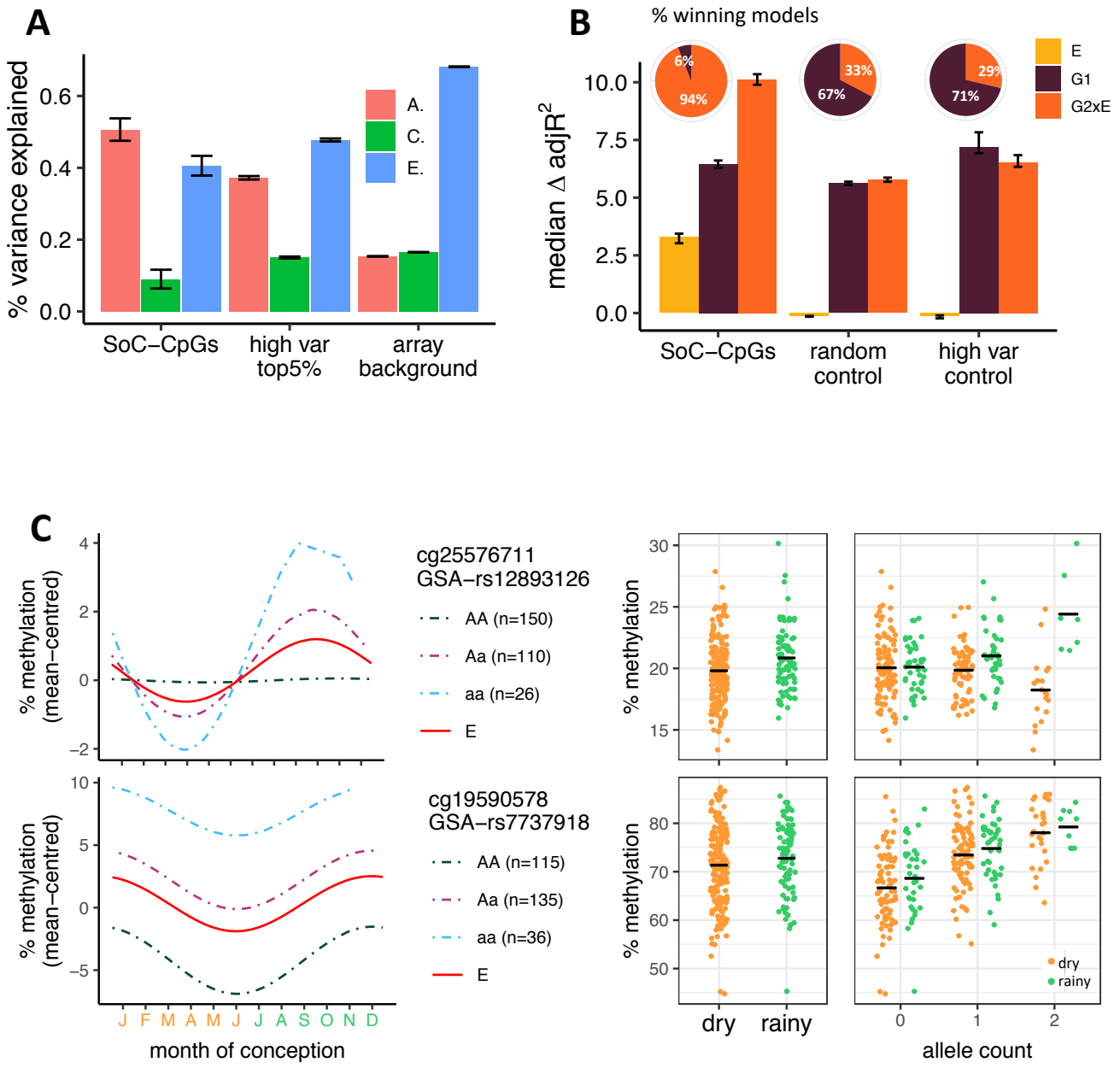

Figure 4. Influence of genotype, periconceptional environment and gene-environment interactions on DNAm. A: Mean methylation variance explained attributable to additive genetic (A), common (C) and non-shared (E) environment effects for replicated CpGs, CpGs in the top 5% by variance and Illumina array background. Estimates for CpGs on the Illumina 450k are from Hannon et al[30]. Error bars represent bootstrapped 95% confidence intervals. **B:** Proportion of methylation variance explained by E, G and GxE models for SoC-CpGs and random and high variance control CpGs. Δ adjR^2 is the additional variance explained by the specified model, over and above a covariate-only model (see Methods for further details). Pie charts show proportion of winning models, assessed using AIC. Note that E-only is never the winning model. **C:** (Left) Examples of ME-CpGs in the replicated set 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, 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.