Environmentally sensitive hotspots in the methylome of the early human embryo

Matt J. Silver^{1*}, Ayden Saffari¹, Noah J. Kessler², Giriraj R. Chandak³, Caroline H.D. Fall⁴, Prachand Issarapu³, Akshay Dedaniya³, Modupeh Betts¹, Sophie E. Moore^{1,5}, Michael N. Routledge^{6,7}, Zdenko Herceg⁸, Cyrille Cuenin⁸, Maria Derakhshan¹, Philip T. James¹, David Monk^{9,10} and Andrew M. Prentice¹

¹Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, UK & Gambia.

²Department of Genetics, University of Cambridge, UK.

³Genomic Research on Complex Diseases (GRC Group), CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India.

⁴MRC Lifecourse Epidemiology Unit, University of Southampton, Southampton General Hospital, UK.

⁵Department of Women and Children's Health, King's College London, UK.

⁶School of Medicine, University of Leeds, UK

⁷School of Food and Biological Engineering, Jiangsu University, Zhenjiang, China

⁸Epigenomics and Mechanisms Branch, International Agency for Research on Cancer, Lyon, France

⁹Biomedical Research Centre, University of East Anglia, UK.

¹⁰Bellvitge Institute for Biomedical Research, Spain.

*Corresponding author: matt.silver@lshtm.ac.uk

Abstract

In humans, DNA methylation marks inherited from gametes are largely erased following fertilisation, prior to construction of the embryonic methylome. Exploiting a natural experiment of seasonal variation including changes in diet and nutritional status in rural Gambia, we analysed three datasets covering two independent child cohorts and identified 259 CpGs showing consistent associations between season of conception (SoC) and DNA methylation. SoC effects were most apparent in early infancy, with evidence of attenuation by mid-childhood. SoC-associated CpGs were enriched for metastable epialleles, parent-of-origin specific methylation and germline DMRs, supporting a periconceptional environmental influence. Many SoC-associated CpGs overlapped enhancers or sites of active transcription in H1 ESCs and fetal tissues. Half were influenced but not determined by measured genetic variants that were independent of SoC. Environmental 'hotspots' providing a record of environmental influence at periconception constitute a valuable resource for investigating epigenetic mechanisms linking early exposures to lifelong health and disease.

Introduction

DNA methylation (DNAm) plays an important role in a diverse range of epigenetically regulated processes in mammals including cell differentiation, X-chromosome inactivation, genomic imprinting and the silencing of transposable elements¹. 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}.

The human methylome is 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 beyond⁴. 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 reprogramming⁵.

The effects of early exposures on the mammalian methylome have been widely studied in animal models and periconceptional and early gestational factors including maternal folate and exposure to famine have been associated with DNAm changes in humans⁶⁻⁸. However, causal pathways are difficult to elucidate in human observational studies, and even randomised experimental designs are prone to confounding due to reverse causation from exposure-related effects⁹.

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^{10,11}. We assess the influence of seasonality on DNAm in two Gambian child cohorts^{12,13}, one with longitudinal data, enabling robust identification of loci showing consistent effects at the ages of 2 years and in mid-childhood (Fig. 1). Through prospective study designs we capture conceptions throughout the year and, in contrast to previous analyses in this population¹⁴⁻¹⁶, we 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 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)^{17,18} and/or epigenetic supersimilarity (ESS)¹⁸ (Fig. 1). These loci bear the hallmarks of metastable epialleles (MEs), loci with methylation states that vary between individuals 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 begins¹⁹, and several MEs have been shown to be sensitive to periconceptional nutrition in mice²⁰. 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 period²¹. We also investigate links with transposable elements and transcription factors associated with the establishment of methylation states in the early embryo, and 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.

Our identification of hotspots in the postnatal methylome that retain a record of environmental conditions during gametic maturation and/or in the very early embryo, provides a valuable resource for the investigation of epigenetic mechanisms linking early-life nutritional and other exposures to lifelong health and disease.

86 Results

87 *Identification of Gambian season of conception associated CpGs*

88 Key characteristics of DNAm datasets from the two Gambian cohorts analysed in this study are
89 provided in Table 1. DNAm differences associated with season of conception are potentially
90 confounded by season of sample collection effects in the ENID 2yr dataset (n=233) since all
91 samples were collected at age 24 months (Fig. 2A top). This is not the case in the older EMPHASIS
92 cohort (n=289; age 7-9 yr) where all samples were collected in the Gambian dry season (Fig. 2A
93 bottom). To account for the potential influence of season of collection effects, we therefore
94 compared year-round DNAm signatures across ENID (2yr) and EMPHASIS datasets by focussing
95 on 391,814 autosomal CpGs ('array background') intersecting the Illumina HM450 and EPIC
96 arrays used to measure DNAm in each dataset (Table 2). We modelled the effect of date of
97 conception on DNAm using Fourier (or 'cosinor') regression²² which makes no prior assumptions
98 about specific seasonal windows that might drive DNAm changes in offspring (see Methods).

99 We began by identifying 768 SoC-associated CpGs showing significant seasonal variation in 2-year
100 olds from the ENID cohort with a false discovery rate (FDR)<5% (Supplementary File 1a;
101 Methods). Fourier regression models revealed a heterogeneous distribution of year-round
102 methylation peaks and nadirs at these loci (Fig. 2B). SoC-associated loci were highly enriched for
103 loci exhibiting SIV/ESS previously identified in multi-tissue screens in adult Caucasians^{17,18},
104 hereafter named 'MEs' for short (Table 3; Fig. 2B, 26 ME CpGs marked in red; enrichment $p =$
105 2.5×10^{-14}). More than twice as many of these loci were within 100bp of a putative ME (n=56;
106 Supplementary File 1a). All identified loci showed increased seasonal amplitudes, defined as the
107 distance between methylation peak and nadir, compared to matched and random controls
108 (Supplementary File 1b; see Table 2 and Methods for justification and further details on selection of
109 controls). Loci with the largest amplitudes tended to show increased methylation in conceptions in
110 the Gambian rainy season (Fig. 2D left) in line with our previous observations in this
111 population^{14,15}.

112 Next, we analysed SoC effects at these 768 loci in 7-9-year olds from the EMPHASIS cohort. Mean
113 methylation at SoC-associated loci was strongly correlated across cohorts (Appendix 1-Fig. 1), and
114 we found evidence of a similar effect of increased methylation in conceptions in the Gambian rainy
115 season (Fig. 2D right). 61 loci (including three ME CpGs) were also significantly associated with
116 SoC (FDR<5%) in the older cohort (Fig. 2D right). Notably, the date of conception at methylation
117 maximum was highly correlated across these two independent and different aged cohorts (Fig. 2C
118 left; Spearman $\rho=0.7$, $p=1.0 \times 10^{-8}$). No significant correlation was observed at matched controls
119 with similar methylation distributions to SoC-associated loci (Fig. 2C right).

120 Given the strikingly similar seasonal patterns across the two cohorts, we next investigated reasons
121 for the smaller number of SoC-associated loci (FDR<5%) in the older (EMPHASIS) cohort.
122 Focussing on the 768 SoC-associated loci in the ENID cohort, we found significant and relatively
123 large effect size (SoC amplitude) decreases in the older cohort, with SoC-effect attenuation most
124 marked at loci that are also putative MEs (4.1% median methylation decrease, Wilcoxon $p=2.8 \times 10^{-6}$;
125 non-ME CpGs: 1.9%, Wilcoxon $p=1.5 \times 10^{-77}$; Fig. 2E; Supplementary File 1c). Corresponding
126 SoC amplitude changes in matched and random controls were much smaller (0.2%, $p=2.0 \times 10^{-8}$ and
127 0.05%, $p=0.02$ respectively; Supplementary File 1c).

128 Noting that loci with larger SoC amplitudes showed a more consistent pattern of seasonal variation,
129 both within and between cohorts (Figs. 2C & 2D), and reasoning that even transient molecular
130 biomarkers of periconceptional environment in postnatal tissues could have biological significance,
131 we focussed on 259 SoC 'hotspots' or SoC-CpGs with FDR<5% and SoC amplitude $\geq 4\%$
132 identified in the ENID 2yr cohort (Fig. 2D left, loci to the right of the dashed line; Table 2;
133 Supplementary File 1d). Sensitivity analysis confirmed that SoC effect estimates at SoC-CpGs were
134 robust to different modelling strategies with respect to estimated blood cell composition and batch
135 effects (see Methods).

136 We tested the hypothesis that SoC amplitudes at SoC-CpGs decrease with age by generating EPIC
137 array methylation data on blood samples in a subset of n=138 individuals from the ENID cohort at
138 5-7yrs (Table 1). This revealed a strongly consistent methylation signature at n=157 replicating
139 SoC-CpGs (Fig. 3A; Supplementary File 1d; Spearman $\rho=0.7$, $p<2.2\times10^{-16}$). This analysis also
140 provided further evidence of SoC effect attenuation with age at SoC-CpGs (Fig. 3B; Supplementary
141 File 1d; Wilcoxon signed rank test for difference in SoC amplitude $p=6.7\times10^{-12}$ for SoC-CpGs and
142 $p=1.0$ for matched controls). ENID longitudinal samples were collected in the rainy season
143 (Appendix 1-Fig. 2), again exhibiting a different confounding structure with respect to season of
144 sample collection compared to samples analysed at the earlier timepoint (ENID 2yr) and in
145 EMPHASIS (Fig. 2A).

146 Critically, the date of conception at methylation maximum at SoC-CpGs was highly consistent
147 across all three datasets analysed, with a distinct pattern of methylation maxima for conceptions
148 falling within the August-September period, most markedly at putative MEs with independent
149 evidence of establishment in the early embryo (Fig. 4A). The August to September period
150 corresponds to the peak of the Gambian rainy season, a strong validation of our previous studies in
151 babies and infants that focussed on conceptions at peak seasons only, with similar observations of
152 higher methylation in conceptions at the peak of the Gambian rainy season compared to peak dry
153 season^{14-16,18}. Methylation minima fall within the February-April period, corresponding to the peak
154 of the dry season (Appendix 1-Fig. 3).

155 Our observation of a remarkably similar season of conception signature across two cohorts and
156 three datasets with different confounding structures with respect to season of sample collection,
157 batch and biological variables (Supp Tables 16-18), combined with evidence from cross-cohort and
158 longitudinal analyses of SoC-effect attenuation with age led us to conclude that SoC-CpGs act as
159 robust sentinels of SoC-associated effects persisting at least until the age of 2 years.

160

161 *Properties of SoC-CpGs*

162 SoC-CpGs are distributed throughout the genome and cluster together in several regions (Appendix
163 1-Fig. 4). Noting that the number of clusters is relatively insensitive to the inter-CpG distance used
164 to define them (Appendix 1-Fig. 5), we identified 56 distinct SoC-CpG clusters and 105
165 ‘singletons’ (SoC-CpGs with no close neighbours) using a maximum inter-CpG distance of 5kbp
166 (Supplementary File 1e). With this definition, 59% of SoC-CpGs fell within clusters
167 (Supplementary File 1f). Of note, SoC effect amplitudes and cross-cohort correlations were greater
168 at SoC-CpGs falling within clusters than with singletons (Appendix 1-Fig. 6).

169 Several SoC-CpG clusters extend over more than 500bp, notably a cluster mapping to *IGF1R* which
170 spans 872bp and covers 7 CpGs (Supplementary File 1e, Appendix 1-Fig. 7). All but one of these 7
171 CpGs were also significantly associated with SoC (FDR<5%) in the older EMPHASIS cohort
172 (Supplementary File 1d).

173 Compared to array background, SoC-CpGs are intermediately methylated, most notably at putative
174 MEs (Fig. 4B), and they tend to fall within CpG islands compared to array background and matched
175 controls (Fig. 4C). SoC-CpGs are also highly enriched for MEs (21-fold enrichment, $p=3.0\times10^{-23}$,
176 cluster-adjusted: 17-fold, $p=3.1\times10^{-11}$; Supplementary File 1g; see Methods for further details on
177 cluster-based adjustments). The number of SoC-CpGs directly overlapping previously identified
178 MEs is small (n=24), although 49 SoC-CpGs (19%) fall within 100bp of a ME (Supplementary File
179 1d). Further investigation revealed that a large majority (n=19/24) of overlapping MEs were
180 identified in our previous screen for SIV, with a smaller number exhibiting ESS (n=7/24;
181 Supplementary File 1g). No MEs overlap methylation distribution-matched controls.

182 Pairwise methylation states are highly correlated at a large majority of SoC-CpGs in both cohorts,
183 so that the same individuals tend to have relatively high or low methylation at multiple SoC-CpGs
184 (Fig. 4D). Pairwise correlations are not driven by increased correlation within SoC-CpG clusters

(Appendix 1-Fig. 8), and methylation at matched and random controls is uncorrelated, thus reducing the possibility that these correlations are driven by statistical artefacts.

Finally, SoC-CpG reliability is classified as ‘excellent’ (median ICC=0.76) using probe reliability estimates from a recent repeated measures study²³. Reliability of matched control CpGs is classified as ‘good’ (median ICC=0.68) using the same method.

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 screen¹⁷. We aligned our data with public reduced representation bisulfite-seq (RRBS) data from human IVF embryos⁴ and obtained informative methylation calls for 112,380 array background CpGs covered at $\geq 10\times$ read depth in inner cell mass (ICM, pre-gastrulation) and/or embryonic liver (post-gastrulation) tissues. As previously noted at putative MEs¹⁷, we found a distinctive pattern of increased incidence of intermediate methylation states at SoC-CpGs in post-gastrulation embryonic liver tissue, strongly contrasting with a general trend of genome-wide hyper- and hypo-methylation at loci mapping to array background (Fig. 5A). A similar pattern of increased incidence of intermediate methylation states was observed at distribution-matched controls.

We previously observed consistent hypomethylation at MEs across all gametic and early embryonic developmental stages, most notably in sperm¹⁷. We tested the latter observation at SoC-CpGs by aligning our data with public sperm WGBS data²⁴, restricting our analysis to 294,240 CpGs mapping to array background that were covered at $\geq 10\times$. SoC-CpGs tended to be hypomethylated in sperm, compared to loci mapping to matched control CpGs and array background respectively (Figs. 5B & 5C). Intermediate methylation states at SoC-CpGs were preserved in both Gambian cohorts irrespective of sperm methylation states, in contrast to array background CpGs where methylation distributions strongly reflected sperm hypomethylation status (ENID cohort: Fig. 4D left; EMPHASIS cohort Appendix 1-Fig. 9 left).

Our observation of an increased incidence of sperm hypomethylation at SoC-associated loci, together with existing evidence that imprinted genes may be sensitive to prenatal exposures^{16,25,26}, prompted us to investigate a potential link between SoC-CpGs and parent-of-origin specific methylation (PofOm). A recent study used phased WGBS methylomes to identify regions of PofOm in 200 Icelanders²⁷. We analysed 699 of these PofOm CpGs overlapping array background (Table 3) and observed strong enrichment for PofOm CpGs at SoC-CpGs and at all MEs on the array (18- and 15-fold enrichment, $p=3.0\times 10^{-8}$ and 1.8×10^{-36} respectively; Supplementary File 1h; Fig. 5D right, PofOm CpGs marked as green triangles). No enrichment was observed at distribution-matched controls (Supplementary File 1h). PofOm enrichment at SoC-CpGs is driven by a large PofOm region spanning 6 CpGs on chr15 at *IGF1R* (Supplementary File 1d); along with two singleton PofOm SoC-CpGs, one on chr18 close to *PAR6G*, and the other in the Prader-Willi syndrome-associated imprinted region neighbouring *MAGE2*, also on chr15. All of these loci have increased methylation in the rainy season, with SoC effect sizes (methylation amplitudes) ranging from 4.1-8.4% (median 6.1%; Supplementary File 1d).

Regions of PofOm detected in postnatal samples tend to be differentially methylated in gametes²⁷, and may thus have evaded epigenetic reprogramming in the pre-implantation embryo²⁵. 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 beyond²⁸. 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 strong enrichment for oocyte (maternally methylated), but not sperm gDMRs, at all PofOm loci identified by Zink *et al*²⁷ (Supplementary File 1h), confirming previous observations of an excess of PofOm loci that are methylated in oocytes only²⁷. This enrichment was particularly

235 strong for oocyte gDMRs (oo-gDMRs) persisting in placenta (Supplementary File 1h). We next
 236 analysed SoC-CpGs and MEs and again found evidence for strong enrichment of oocyte, but not
 237 sperm gDMRs at these loci (6.2-fold oo-gDMR enrichment, $p=2.3 \times 10^{-16}$ at SoC-CpGs; 2.9-fold,
 238 $p=1.2 \times 10^{-24}$ at MEs), including after adjustment for CpG clustering (Supplementary File 1h). Of
 239 note, 14% (36/259) of SoC-CpGs overlapped oo-gDMRs, in strong contrast to matched and random
 240 controls (Fig. 5C). These clustered into 19 distinct oo-gDMR regions (Supplementary File 1i) -
 241 more than six times the number identified as exhibiting PofOm by Zink *et al.* (3 regions;
 242 Supplementary File 1d).

243 A large majority of SoC-CpGs that are hypomethylated in sperm are not oo-gDMRs (i.e. they are
 244 not hypermethylated in oocytes) (Fig. 5C & 5D bottom right), suggesting that factors associated
 245 with regional sperm hypomethylation rather than differential gametic methylation may be a key
 246 driver of sensitivity to periconceptional environment at these loci.

247 *SoC-CpG overlap with predicted chromatin states*

248 We assessed the overlap of SoC-CpGs with predicted chromatin states generated from histone
 249 marks in various cell lines and tissues by the Roadmap Epigenomics Consortium²⁹. Given our
 250 interest in methylation states associated with periconceptional environment that persist into early
 251 postnatal life, we focussed on data from H1 embryonic stem cells (ESCs) and fetal tissues (fetal
 252 brain, muscle and small intestine) derived from all 3 germ layers, described as having the ‘highest
 253 quality’ epigenomes (see Fig. 2 in Kundaje *et al.*²⁹). Around half of all SoC-CpGs overlapped sites
 254 with predicted transcriptional or regulatory function, with relatively few overlapping constitutive
 255 heterochromatic regions (Fig. 6). Overlaps with specific histone marks in H1 ESCs are given in
 256 Appendix 1-Fig. 10. As expected, given predicted chromatin states, many show a predominance of
 257 overlapping H3K4me1 and H3K27me3 marks and combinations thereof, suggestive of active or
 258 poised enhancers.

259 *Enrichment of transposable elements and transcription factors associated with genomic imprinting*

260 Variable methylation states at MEs are associated with transposable elements (TEs) in murine
 261 models^{30,31}, and we have previously observed enrichment for proximity to two classes of
 262 endogenous retroviruses, ERV1 and ERVK, at putative human MEs^{16,17}. Here we found evidence
 263 that SoC-CpGs are enriched for proximity to ERV1 (Fig. 7A top) but not ERVK retroviral elements
 264 (Appendix 1-Fig. 11; Table 3).

265 Enrichment for PofOm and gDMRs at SoC-CpGs suggests a potential link to mechanisms
 266 implicated in the maintenance of PofOm and genomic imprinting in the early embryo. Our previous
 267 analysis of MEs identified from WGBS data found enrichment for proximal binding sites for 3
 268 transcription factors (TFs: CTCF, ZFP57 and TRIM28) identified through ChIP-seq of embryonic
 269 stem and kidney cells that are linked to maintenance of PofOm at imprints¹⁷. Here we found
 270 evidence for enrichment of proximal ZFP57 binding sites within 2kbp of a SoC-CpG as previously
 271 observed at MEs, but we found no evidence for enrichment of proximal CTCF or TRIM28 binding
 272 sites in this array-based study (Fig. 7A bottom; Appendix 1-Fig. 11; Table 3).

273 *Influence of genotype*

274 Genetic variation is a major driver of inter-individual variation in DNAm via methylation
 275 quantitative trait loci (mQTL)³². We explored the influence of mQTL on SoC-CpGs in the
 276 EMPHASIS cohort for which we had genotype data on 284 individuals measured at >2.6M SNPs
 277 after imputation from the Illumina Global Screening Array (GSA) and subsequent LD-pruning (see
 278 Methods). The majority of mQTL effects occur in *cis*³². In order to maximise power we therefore
 279 adopted a two-step approach where we performed separate screens for mQTLs in *cis* (defined as
 280 SNPs within 1Mb of an associated CpG³²) and *trans* (all others), and compared our findings at SoC-
 281 CpGs with matched and random control CpGs (Methods). Half of SoC-CpGs had one or more
 282 associated mQTL compared with 78% and 19% of matched and random controls respectively. 92%
 283 of SoC-CpG mQTL were in *cis* (Table 4).

284 We next compared methylation variance explained by significant mQTL using adjusted R^2 values
285 for all SoC-CpGs and controls with at least one genome wide significant mQTL ($FDR < 5\%$; $n=130$,
286 201 and 50 CpGs for SoC-CpGs, matched and random controls respectively; Table 4). These values
287 were compared to a baseline model that included the same set of covariates (principal components,
288 age and sex) used in Fourier regression models for the main seasonality analysis, in order to account
289 for potential differences in additional variance explained by other covariates and unmeasured
290 factors (see Methods). There was no difference in additional variance explained by significant
291 mQTL between SoC-CpGs and both sets of control CpGs (Fig. 7B; Table 4).

292 To assess the potential for genetic confounding of SoC-associated DNAm signals at SoC-CpGs
293 with associated mQTL, we tested all SoC-CpG-mQTL for association with season of conception
294 using an allelic model. After accounting for multiple testing, no significant SoC-mQTL associations
295 were identified (Supplementary File 1j and Methods). Our observations that i) SoC-CpGs are
296 distributed throughout the genome; ii) SoC-CpG mQTL occur primarily in *cis*; and iii) none are
297 associated with season of conception; strongly suggest that SoC-methylation associations at SoC-
298 CpGs are not confounded by genetic variation.

299 Finally, we searched for gene-environment (SoC) interaction (GxE) effects, again performing
300 separate tests for SNPs in *cis* and *trans*. No GxE associations were identified after correcting for
301 multiple testing (see Methods).

302 *Influence of genetic ancestry*

303 Eighty percent of the population of the Kiang West region of The Gambia from which the cohorts
304 analysed in this study are drawn are of Mandinka ethnicity, with the majority of the remainder
305 Fula³³. This is evident from a genome-wide principal component analysis (PCA) of genetic
306 variation in the EMPHASIS cohort, where we observed a distinct cluster of 16 individuals from a
307 single village which is predominantly Fula (Appendix 1-Fig. 12). Individuals in the main ENID
308 cohort were drawn from the same Kiang West villages as the EMPHASIS study, but we were
309 unable to directly adjust for potential confounding effects due to genetic ancestry since no genetic
310 data was available for this cohort. Based upon the EMPHASIS cohort PCA and our knowledge of
311 village population structures, we reasoned that village of origin is a useful proxy for genetic
312 ancestry in the ENID cohort and performed a sensitivity analysis with an additional covariate
313 dichotomised according to whether an individual came from the predominantly Fula village. The
314 first two genetic principal components were used as adjustment covariates for the corresponding
315 EMPHASIS analysis. Results from this ethnicity-adjusted sensitivity analysis were not materially
316 different from those obtained for the main analysis (Appendix 1-Fig. 13; Supplementary File 1s).

317 *Overlap of SoC-CpGs with existing studies*

318 To place our findings in the context of existing literature on associations between DNAm and
319 nutrition-related exposures, including exposure to famine conditions, folate supplementation in
320 gestation and previous reported associations with Gambian SoC, we checked for overlaps between
321 SoC-CpGs and loci identified in a recent review by James et al²⁶. Many cited studies including the
322 majority of previous work in The Gambia used pyrosequencing and other methylation platforms
323 targeting loci not covered by Illumina arrays. However, a total of 57 previously identified loci did
324 overlap or partially overlap array background. None of these overlapped a SoC-CpG within 1kbp.
325 We also checked for overlaps with the larger set of SoC-associated CpGs not passing the 4%
326 minimum effect size threshold, and found a single CpG (cg17434309) mapping to *IGF2* that was
327 within 1kbp of two previously identified loci, one linking maternal plasma vitamin B12 with cord
328 blood methylation³⁴, and the second linking gestational famine to blood methylation in older
329 adults³⁵.

330 We next looked for overlaps between SoC-CpGs and CpGs identified in the EWAS Catalog
331 (<http://ewascatalog.org/>), a manually curated database of significant results ($p < 1 \times 10^{-4}$) from
332 published epigenome-wide association studies (EWAS). This search produced published
333 associations for 167 out of the 259 SoC-CpGs, mapping to 27 unique traits covering a range of pre-

334 and post-natal exposures (Supplementary File 1l). Noteworthy amongst the most frequently
335 reported associations with SoC-CpGs (Supplementary File 1m) in the context of our study were
336 those with sex, gestational age, maternal smoking in pregnancy, maternal plasma folate levels and
337 adult body mass index (BMI) with 109, 45, 16, 6 and 1 associated SoC-CpG(s) respectively.

338 We investigated some of these links using data from the 2yr ENID cohort considered in our main
339 analysis and confirmed multiple significant associations with infant sex but not gestational age or
340 maternal folate at conception (Supplementary File 1n). Links with adult BMI and maternal smoking
341 were not considered as adult BMI was not available and the incidence of smoking is extremely low
342 in our study population.

343 All Fourier regression models in our main season of conception analysis included sex as an
344 adjustment covariate. The finding that multiple SoC-CpGs were associated with sex in the EWAS
345 Catalog, with replication of this association in the ENID cohort, was therefore surprising. This
346 prompted us to check for a residual confounding effect due to sex by repeating our analysis with
347 methylation values pre-adjusted for sex using a regression model with sex as the only adjustment
348 covariate, prior to running the full regression models. This produced near identical results to the
349 main analysis without pre-adjustment. This, combined with our observation that sex is not
350 associated with any tested batch or biological covariates (Supplementary Files 1p-1r), strongly
351 suggests that the observed SoC associations were not driven by confounding due to sex.

352 Finally we searched for SoC-CpGs within 1kbp of SNP associations ($p < 1 \times 10^{-5}$) in the GWAS
353 Catalog³⁶, since DNAm could mediate GWAS signals in genomic regions where functional effects
354 are difficult to elucidate³⁷. 11 SoC-CpGs mapped to a total of 12 SNPs associated with 8 unique
355 traits in the GWAS Catalog (Supplementary File 1o). Notable traits from a developmental
356 programming perspective were those linked to childhood obesity³⁸ (2 SoC-CpGs) and QRS traits
357 associated with cardiovascular mortality in adults (1 SoC-CpG)^{39,40}.

358 Discussion

360 We have utilised a natural, seasonal experiment in rural Gambia whereby human conceptions are
361 'randomised' to contrasting environmental (especially dietary) conditions to examine whether this
362 differential exposure leaves a discernible signature on the offspring methylome. We analysed
363 methylation data from two independent, different-aged cohorts and identified 259 'SoC-CpGs' with
364 evidence of sensitivity to season of conception (SoC) in infancy. We found evidence of SoC effect
365 attenuation with age, but SoC-related temporal patterns were nonetheless strikingly similar,
366 suggestive of a common effect of periconceptional environment. Importantly, the analysed datasets
367 have contrasting confounding structures, notably with regard to the timing of sample collection, the
368 latter eliminating potential confounding due to seasonal differences in leukocyte composition.
369 These results, derived from analysis of Illumina array data with limited coverage, suggest there may
370 be many more hotspots sensitive to the periconceptional environment across the human methylome.

371 This analysis builds on previous epigenetic studies in this setting that have focussed on single
372 cohorts and analysed methylation differences between individuals conceived at the peaks of the
373 Gambian dry and rainy seasons only^{14-16,18,41}. Increased methylation in offspring conceived at the
374 peak of the Gambian rainy season is consistent with previous findings and this observation is now
375 greatly strengthened by the application of Fourier regression to model year-round conceptions, an
376 approach that makes no prior assumption of when methylation peaks and nadirs may occur. The
377 number of identified SoC-CpGs is also substantially increased in this study. Comparisons with
378 array background and control CpGs matching SoC-CpG methylation distributions increase
379 confidence that these findings are not statistical artefacts.

380 Triangulation with other public data provides multiple lines of evidence supporting the notion that
381 methylation states at these loci are established in the periconceptional period. First, they are highly
382 enriched for putative MEs and related loci identified in other studies with characteristic methylation
383 signatures suggestive of establishment early in embryonic development^{17,18}. Second, like MEs,
384 season-associated loci exhibit unusual methylation dynamics in early stage embryos¹⁷. Third, they

385 have distinctive gametic methylation patterns, notably hypomethylation in sperm (in common with
386 putative MEs¹⁷), and differential gametic and parent-of-origin specific methylation in a subset.
387 Increased incidence of hypomethylated states in sperm at SoC-CpGs may reflect their enrichment at
388 CpG islands⁴², sequence features that are largely refractory to protamine exchange, with the
389 possibility for retaining epigenetic function associated with histone modifications into the early
390 embryo⁴³. Fourth, many overlap H3K4me1 and H3K27me3 marks which coordinate transient gene
391 expression and early lineage commitment in human ESCs, in part through demarking poised
392 enhancers⁴⁴.

393 A large majority of SoC-CpGs have not previously been identified as MEs, but given the supporting
394 evidence described above, we speculate that many are likely to be so. Indeed, evidence of
395 attenuation of SoC effects with age suggests that, to the extent that interindividual variation is
396 driven by periconceptional environmental factors, screens for putative MEs (including ESS and
397 SIV) in adult tissues used as a reference in this analysis may be missing metastable regions which
398 are more pronounced earlier in the life course. Evidence of much larger SoC effect sizes at known
399 MEs in both Gambian cohorts supports this. Furthermore, identification of putative human SIV-
400 MEs requires analysis of datasets with tissues derived from different germ layers for each
401 individual. Such datasets are expensive to acquire and rare, meaning that the few published ME
402 screens are likely to be underpowered due to small sample sizes.

403 The observed attenuation of SoC effects with age has implications for detecting the effect of
404 periconceptional exposures on DNAm in samples collected beyond the neonatal and early
405 childhood periods, an important consideration for epigenetic epidemiological studies since non-
406 persisting methylation differences could still have a significant impact on early developmental
407 trajectories with life-long consequences^{45,46}. The methylome changes with age, a phenomenon
408 sometimes referred to as ‘epigenetic drift’⁴⁷. Attenuation of SoC effects could be a consequence of
409 localised or global epigenetic drift or other processes such as clonal selection⁴⁸.

410 Interestingly, more SoC-CpGs replicated in the smaller ENID (5-7yr) dataset than in the older
411 EMPHASIS (7-9yr) cohort (n=157 vs 61 respectively). This could be due to age or other cohort-
412 specific differences, or it might reflect differences in seasonality such as secular changes in the
413 intensity and duration of the rainy season.

414 Intra-individual methylation states at SoC-associated loci are highly correlated across loci despite
415 being distributed throughout the genome, suggesting that a common mechanism is at play. MEs
416 located within intracisternal A particle (IAP) transposable elements have been the focus of many
417 studies of SIV in mice, the *Agouti* viable yellow locus being a paradigm example. While one study
418 found that methylation does not co-vary across different murine IAP MEs in the absence of
419 environmental perturbations³¹, another recent study identified several IAP MEs located on different
420 chromosomes which are modified by the same cluster of KRAB zinc finger proteins⁴⁹,
421 demonstrating that genetic mechanisms for the simultaneous epigenetic regulation of multiple loci
422 do exist.

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

426 First, we observed strong enrichment for germline differentially methylated regions, with 14% of
427 SoC-CpGs overlapping 19 gDMRs hypomethylated in sperm and hypermethylated in oocytes. A
428 minority of these show evidence of PofOm persisting in postnatal tissues. This observation aligns
429 with a growing body of evidence linking early environment, notably nutritional factors involved in
430 one-carbon (C1) metabolism, with methylation at imprinted regions^{25,26}. Indeed we have previously
431 noted an association between season of conception and several C1 metabolites at a maternally
432 imprinted region at the small non-coding RNA *VTRNA2-1*¹⁶, consistent with evidence of
433 ‘polymorphic imprinting’ linked to prenatal environment at this locus^{27,50}. Furthermore, we
434 previously found strong enrichment for proximal binding sites of several transcription factors (TFs)
435 associated with the maintenance of PofOm in the early embryo at MEs detected in a WGBS

436 screen¹⁷. We were only able to replicate enrichment for one of these, ZFP57, at SoC-CpGs in this
437 study. This may reflect the relatively small proportion of PofOm loci in the set of SoC-CpGs, or
438 factors related to the limited methylome coverage of Illumina arrays. More targeted experimental
439 work is required to determine the extent of SoC effects at imprinted loci, especially given our
440 observation that SoC-CpGs are often proximal to ERV transposable elements that have recently
441 been shown to drive the establishment of germline-derived maternal PofOm⁵¹. Hotspots with
442 evidence of PofOm could be driven by an environmentally sensitive gain of methylation on the
443 paternal allele that is propagated through development, incomplete reprogramming on the maternal
444 allele leaving residual traces of methylated cytosines, or modest *de novo* methylation at some later
445 point. A deeper understanding of mechanisms will require further investigation in cell and animal
446 models.

447 Second, our observation of enrichment for proximity to ERV1 transposable elements at SoC-CpGs
448 aligns with our previous finding at MEs^{16,17}, and is notable since most environment-sensitive mouse
449 MEs are associated with IAPs (which are rodent-specific ERVs)³¹. KRAB zinc-finger protein
450 (KZFP)-mediated repression of transposable elements (TEs) including ERVs has also been
451 proposed as a driver of the rapid evolution of gene regulation⁵². The KZFP ZFP57 is particularly
452 interesting in this respect since its binding to DNA is linked both to repression of TEs and to the
453 maintenance of genomic imprints in the pre-implantation embryo^{53,54}. We previously identified a
454 putative SoC-associated DMR in the *ZFP57* promoter in blood from younger Gambian infants
455 (mean age 3.6 months)¹⁶. It is possible that non-replication of the SoC-association at *ZFP57* in the
456 older Gambian cohort (and of the *VTRNA2-1* SoC-DMR mentioned above which was also identified
457 in younger infants) reflects the more general attenuation of SoC effects described above.
458 Interestingly there is some evidence that the *ZFP57* DMR, which lies 3kb upstream of the
459 transcription start site, is established in the early embryo^{18,55}, and that DNAm at this locus measured
460 in neonatal blood is associated with maternal folate levels^{6,56,57}. Given the important function of
461 ZFP57 in pre-implantation methylation dynamics, its potential role as an environmentally sensitive
462 regulator of multi-locus DNAm effects remains an open question.

463 Third, DNAm at SoC-CpGs is enriched for intermediate methylation states. Intermediate
464 methylation has also been observed at MEs in Gambians and in non-Africans^{8,14,15,41}, and this
465 coincides with a similar observation at MEs in post-gastrulation embryonic tissues¹⁷. Intermediate
466 methylation appears to be predominantly driven by variegated (intercellular) differences in
467 methylation status within a sampled tissue, rather than by allele-specific methylation⁵⁸. Observed
468 differences in aggregate methylation at SoC-CpGs could thus reflect a direct influence of
469 periconceptional environment on the establishment and maintenance of DNAm states in the early
470 embryo at the cellular level, or ‘epigenetic selection’ whereby epigenetically distinct cells form a
471 substrate for clonal selection during development, for example as a potential adaptation to
472 differential metabolic exposures⁵⁹. Stronger enrichment for putative MEs exhibiting SIV compared
473 to those identified through ESS supports establishment in the post-implantation embryo, since
474 methylation differences at ESS loci are presumed to originate in the pre-implantation embryo prior
475 to embryo cleavage in MZ twins¹⁸.

476 The largest SoC-CpG cluster with 7 CpGs is on chromosome 15 and falls within the second intron
477 of *IGF1R*. This bears the hallmarks of being a promoter or active/poised enhancer in multiple cell
478 lines (Appendix 1-Fig. 14). Six CpGs in the *IGF1R* cluster have evidence of PofOm and show a
479 relatively large SoC effect size (median SoC methylation amplitude) of 6.1%. Zink *et al*²⁷ were
480 unable to demonstrate PofO allele-specific expression in this region although others have found
481 evidence of maternal imprinting of an intronic lncRNA at this gene in cancerous cells^{60,61}. Loss of
482 IGF1 receptors gives rise to a major decrease in expression at multiple imprinted genes in mice,
483 suggesting a pathway by which *IGF1R* might regulate growth and metabolism during early
484 development⁶². IGF1R signalling is implicated in fetal growth, glucose metabolism and cancer⁶³⁻⁶⁵,
485 and DNAm differences at *IGF1R* have been observed in birthweight-discordant adult twins⁶⁶.
486 Another SoC-CpG with evidence of PofOm, also on chromosome 15, falls within the known
487 Prader-Willi syndrome-associated paternally expressed imprinted region.

Our observation that 109 out of 259 SoC-CpGs have been associated with sex in previous studies is intriguing. Given the relatively small number (compared to array size) of autosomal sex-linked loci identified in large studies on the Illumina 450k array^{67,68}, this represents a very strong enrichment. We replicated a significant sex association at these loci in the ENID cohort analysed in this study. Our regression analyses were adjusted for sex, and additional sensitivity analyses with DNAm pre-adjusted for sex strongly suggest that our main findings are not confounded by sex. Interestingly, sex-associated loci are enriched at imprinted regions and sex-discordance at autosomal CpGs has been linked to androgen exposures *in utero*^{67,69}. There is also evidence of sex differences in methylation at *DNMT3A/B* and *TET1* genes involved in *de novo* methylation and de-methylation pathways^{67,69}, suggesting a possible interaction between sex-linked epigenetic changes and periconceptional environment during reprogramming in the early embryo. A deeper understanding of potential mechanisms underpinning the observed enrichment of sex effects at loci associated with periconceptional environment requires further functional analysis in cell and/or animal models.

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 cohort and found no evidence of SoC-associated genetic variants driving inter-individual methylation differences at SoC-associated loci in *cis* or *trans*.

However, we did find that half of SoC-CpGs had at least one associated mQTL, indicating the presence of independent additive effects of environment and genetics at these loci, as has been suggested previously at other loci sensitive to pre/periconceptional nutritional exposures^{35,70}. 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 populations¹⁷, 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.

Further work is required to investigate the functional relevance of DNAm changes at SoC-CpGs, some of which are relatively small (around 4% SoC amplitude). However, we can speculate that observed methylation changes, which may reflect changes in the chromatin landscape, have the potential to influence gene expression and early development, since our chromatin state analysis predicts that many overlap regions with functional significance in H1 ESCs and fetal tissues. A similarly modest DNAm change at a locus in the *POMC* gene that is associated with SoC in blood has been linked to obesity risk in German children and adults^{41,71}, and to differential transcription factor binding and differences in *POMC* expression⁷¹. Furthermore, blood DNAm at a putative ME within the *PAX8* gene has been linked to SoC in Gambian infants¹⁵ and is associated with thyroid volume and function in Gambian children, and with certain maternal nutrition biomarkers at conception⁷². *PAX8* DNAm is also associated with expression of the anti-sense gene *PAX8-AS1* (alternatively known as *LOC654433*) in thyroid tissue⁷². No loci from either study overlap array background in this study. More generally, DNAm at SIV loci associated with periconceptional environment has been associated with a number of diseases including Alzheimer's, cancer, rheumatoid arthritis and schizophrenia²¹.

We have previously shown that several metabolites involved in one-carbon methylation pathways show significant seasonal variation in maternal blood plasma in this population which is largely dependent on subsistence farming^{11,15}. While we suspect that seasonal differences in nutrition are likely drivers of the effects that we have observed, analysis of the links between maternal nutritional biomarkers and DNAm is challenging due to the complex interdependence of one-

carbon metabolites acting as substrates and cofactors within one-carbon pathways⁷³. Furthermore, it is possible that DNAm differences are linked to other aspects of Gambian seasonality, such as variation in pesticide use or infectious disease burden^{10,74}.

Measured DNAm values at SoC-CpGs and matched controls have previously been reported to be reliable, in strong contrast to the majority of probes overlapping the 450k and EPIC arrays that have been found to have low test-retest reliability²³. This likely reflects increased inter-individual variability and/or intermediate DNAm at SoC-CpGs and controls^{75,76}.

Limitations of this analysis include a lack of genetic data for the ENID cohort, so that our genetic analyses are restricted to the older EMPHASIS cohort where SoC effects are smaller. Furthermore, comparator data on MEs and gamete, embryo and PofO methylation are derived from non-African individuals so we cannot assess the influence of ethnicity on these analyses. We also note that there is some inter-relatedness in this study population which practices polygamy. However, of the 199 individuals with available data on paternal identity in the ENID cohort, 2% have a shared father suggesting that this is unlikely to influence the main findings.

There is increasing interest in the phenomenon of methylation variability as a marker of disease and of prenatal adversity, and in genetic variation as a potential driver of methylation variance⁷⁷. This raises the possibility 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. Our gene-environment interaction analysis likely lacked power to detect gene-environment (SoC) interactions, so that it was not possible to investigate the relative contributions of stochastic, environmental and genetically-mediated variance effects on the establishment of periconceptional SoC-associated methylation states in this study. Nonetheless, the proposition that environmentally sensitive epigenetic signals are selected through their ability to direct phenotypic development to better fit the anticipated future environment, leading to maladaptation and future disease if the environment changes, is intriguing and worthy of further investigation⁵.

Materials and Methods

Gambian cohorts and sample processing

Detailed descriptions of the Gambian cohorts analysed in the season of conception study are published elsewhere^{12,13}. Briefly, for the ENID 2yr dataset, blood samples from 233 children aged 24 months (median[IQR]: 731[729,733] days old) were collected from participants in the **E**arly **N**utrition and **I**mmune **D**evelopment (“ENID”) study¹², born in 2011 and 2012. DNA was extracted, bisulfite-converted and hybridised to Illumina HumanMethylation450 (hereafter “HM450”) arrays following standard protocols (see Van Baak *et al*¹⁸ for further details). N=138 ENID participants with 2yr HM450 data who were enrolled in a follow-up study with DNA collected at 5-7yr (6.2 yrs [5.7,6.6]) had DNA bisulfite-converted and hybridised to Illumina Infinium Methylation EPIC (hereafter “EPIC”) arrays. For the EMPHASIS cohort, DNA was extracted from blood samples from 289 Gambian children aged 7-9yr (9.0 [8.6,9.2] years) participating in the **E**pigenetic **M**echanisms linking **P**re-conceptional nutrition and **H**ealth **A**ssessed in **I**ndia and **S**ub-Saharan Africa (“EMPHASIS”) study¹³, born between 2006 and 2008. DNA was bisulfite-converted and hybridised to 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^{13,78}. In this case date of conception was calculated as date of birth minus 280 days which is the average gestational length for this population.

Samples from the ENID 2yr dataset were processed in two batches with each batch covering conception dates throughout the year. Samples from the ENID (5-7yr) and EMPHASIS (7-9yr) datasets were processed in single batches.

Methylation array pre-processing and normalisation

Raw intensity IDAT files from the HM450 and EPIC arrays were processed using the *meffil*⁷⁹ 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) probes⁸⁰; 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, followed by *Functional Normalisation* to reduce technical variation based on principal component analysis of control probes on the arrays⁸¹. After pre-processing and normalisation, methylation data comprised methylation Beta values for 421,026 CpGs on the HM450 array for the 233 individuals from the ENID 2yr cohort, and 802,283 CpGs on the EPIC array for 289 individuals from the EMPHASIS cohort. 391,814 CpGs intersecting both arrays were carried forward for statistical analysis of these two datasets. Finally, SoC-CpGs, matched and random controls, all of which were present in the ENID (5-7yr) EPIC dataset, were included in the longitudinal analysis.

Statistical modelling

Variation of DNAm with date of conception was modelled using Fourier regression²². 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 value⁸²;

α_{0j} is an intercept term;

α_{ik} is the k^{th} of m adjustment covariates;

θ_i is the date of conception in radians in the interval $[0, 2\pi]$, with 1st January = 0 and 31st

December = 2π , modelled as n pairs of Fourier terms, $\sin \theta_i + \cos \theta_i + \dots + \sin n\theta_i + \cos n\theta_i$;

β_{rj} and γ_{rj} are the estimated regression coefficients for the r^{th} 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 year) and amplitude (distance between methylation 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 ($n=1$) and assessed goodness of fit by comparing full and covariate-only models using likelihood ratio tests. For all datasets, 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 Files 1p-1r). Additional checks confirmed no seasonal variation in estimated white cell composition in the ENID 2yr and EMPHASIS 7-9yr cohort used in the main analysis (see below). 450k Sentrix Column was included as an additional adjustment covariate for the ENID 2yr cohort since this was not robustly captured by any of the first 6 PCs (Supplementary File 1p). Child age was included as an additional adjustment covariate for the ENID 5-7yr and EMPHASIS cohorts, as was maternal nutritional intervention group (see Chandak *et al*¹³ and Saffari *et al*⁷⁰ for further details).

For each CpG j , coefficient estimates β_j , γ_j were determined by fitting regression models 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.

To reduce the influence of methylation outliers, methylation values greater than three times the methylation inter-quartile range (IQR) beyond the 25th or 75th centiles were excluded prior to fitting the models. This resulted in 382,095 (267,369) outliers, or an average of 0.98 (0.68) outliers per CpG being removed from the ENID and EMPHASIS methylation datasets respectively.

Note that our observation of attenuation of SoC effect sizes with age at SoC-CpGs when comparing DNAm data at aged 2yr (ENID) and 7-9yr (EMPHASIS), supported by analysis of longitudinal data (ENID 5-7yr), suggested that a combined meta-analysis of ENID and EMPHASIS datasets would have reduced power to detect SoC effects at 2 years of age.

Sensitivity analyses

We performed sensitivity analyses following the same Fourier regression modelling strategy as outlined above, but with i) methylation estimated cell counts (using *estimateCellCounts()* from *minfi* (v1.30.0)); ii) known batch and technical covariates; and iii) village ID included as additional covariates. See Supplementary File 1s for further details on regression models used in the sensitivity analyses. For each sensitivity analysis we tested to see if Fourier regression coefficients estimated in the main analysis fell within 95% CIs in the adjusted models.

Inflation of test statistics

The concept of genomic inflation rests on the assumption that a relatively small number of loci will be associated with the exposure (or disease/outcome) of interest. Test statistics for the ENID (2yr) cohort did show signs of genomic inflation ($\lambda = 1.33$), suggesting a potential effect of season of conception on global methylation levels⁸³. A similar level of inflation was observed before in a study looking at the effect of periconceptional folate on DNAm⁶. There is also evidence of global and/or multi-locus effects of folate in other studies, including an RCT of folate supplementation in pregnancy⁵⁷, and there are many other examples including studies investigating the effect of mutations in the *MTHFR* gene (see review by Crider et al⁸⁴). While we do not know if the SoC associations we observe in our cohorts are driven by seasonal differences in folate, we do observe significant seasonal differences in multiple one-carbon (C1) metabolites including folate in our population^{11,15}, so that SoC may serve as a proxy for multi-locus C1 metabolite effects in our analyses.

We tested for potential SoC effects on global methylation by analysing methylation differences at LINE1 and Alu elements using REMP (v1.16.0), a recently published method to predict methylation at repetitive elements from Illumina array data⁸³. We used default cutoffs for reliability suggested by the authors. This analysis confirmed a small but significant effect of increased methylation in rainy vs dry conceptions at LINE1 and Alu elements (mean rainy season increase 0.02% and 0.01% respectively, both Wilcoxon $p < 2.2 \times 10^{-16}$) suggesting a SoC effect on global methylation levels. This contrasted with a significant but extremely small effect in the opposite direction across array background (-5×10^{-4} %; $p < 2.2 \times 10^{-16}$; see Appendix 1-Fig. 15), supporting evidence from previous studies of a potential effect of peri/preconception nutrition-related exposures on global methylation levels.

Identification of SoC-CpGs

For the ENID 2yr cohort, 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 with method= 'fdr'. Following the rationale described in the main text, SoC-associated CpGs with a SoC amplitude $< 4\%$ in the ENID cohort were then excluded to form the final set of 259 '**SoC-CpGs**' (see Table 2).

Selection of control CpGs

SoC-CpGs are enriched for intermediate methylation states (Fig. 4B), so that there is a risk that some downstream analyses reflect the distributional properties of these loci, rather than factors associated with their putative establishment at periconception. For this reason we identified a set of '**matched control**' CpGs that were selected to have similar methylation Beta distributions to SoC-CpGs (and additional SoC-associated CpGs with amplitude $< 4\%$) in the ENID 2yr dataset. Matched controls were drawn from array background (excluding SoC-associated CpGs and known MEs/ESS/SIV CpGs), with one matched control identified for each of the 768 SoC-associated CpGs. Alignment of control and SoC-CpG methylation distributions was achieved using a two-sided Kolmogorov-Smirnov test for divergence of cumulative

695 distribution functions (*ks.test()* in R) with a p-value threshold $p > 0.1$. Examples are given in Appendix 1-Fig.
696 16, along with a comparison of sample mean distributions.

697 An additional 768 **random control** CpGs were randomly sampled from array background, again excluding
698 SoC-CpGs and known MEs/ESS/SIV CpGs.

699 *CpG sets considered in analyses*

700 Summary information on external datasets considered in the analyses is provided in Table 3. Further
701 information on these is provided below.

- 702 i. 1,881 *putative ME* CpGs overlapping array background from one or more of the following
703 curated sets of loci: putative MEs exhibiting SIV identified in a multi-tissue WGBS screen in
704 individuals of European and African-American decent described in Kessler *et al*¹⁷; and CpGs
705 exhibiting ‘epigenetic supersimilarity’ and/or SIV derived from analysis of 450k data from
706 individuals of European decent, as described in Van Baak *et al*¹⁸.
- 707 ii. 699 *parent-of-origin-specific CpGs (PofOm CpGs)* overlapping array background from 229
708 regions with PofOm identified in Supplementary Table 1 from Zink *et al*²⁷. PofOm identified in
709 peripheral blood from Icelandic individuals.
- 710 iii. *RRBS early stage embryo data* from Chinese embryos described in Guo *et al*⁴ downloaded
711 from GEO (accession number GSE49828). Only CpGs covered at $\geq 10\times$ in pre-gastrulation inner
712 cell mass and/or post-gastrulation embryonic liver were considered in this analysis. Further
713 details are provided in Kessler *et al*¹⁷.
- 714 iv. *Sperm methylation data* from Japanese donors described in Okae *et al*²⁴ downloaded from the
715 Japanese Genotype-phenotype Archive (accession number S00000000006). Only CpGs covered
716 at $\geq 10\times$ were considered in this analysis.
- 717 v. *Germline gDMRs (gDMRs)*, defined as contiguous 25 CpG regions that were hypomethylated
718 (DNAm mean +1SD < 25%) in one gamete and hypermethylated (DNAm mean-1SD > 75%) in
719 the other, were previously identified from WGBS data by Sanchez-Delgado *et al*²⁸. Persistence
720 of PofOm to the blastocyst and placental stages was established by identifying overlapping
721 intermediately methylated regions in the relevant embryonic tissues, with confirmation of
722 PofOm expression at multiple DMRs²⁸. Japanese and US donors. See Sanchez-Delgado *et al*²⁸
723 for further details.
- 724 vi. *Transposable elements (ERVs)* determined by RepeatMasker were downloaded from the UCSC
725 hg19 annotations repository.
- 726 vii. ZFP57, TRIM28 and CTCF transcription factor binding sites identified from ChIP-seq in human
727 embryonic kidney and human embryonic stem cells used in this analysis are described in Kessler
728 *et al*¹⁷.

729 *Cluster-based adjustments*

730 Many SoC-CpGs cluster together and this could influence some analyses. For example, methylation at CpGs
731 may be highly correlated, which could influence comparisons of inter-CpG correlations between SoC-CpGs
732 and controls (Fig. 3D). Also enrichment tests are likely to be influenced by neighbouring CpGs that together
733 constitute a single ‘enrichment signal’ proximal to a particular genomic feature (e.g. a transcription factor
734 binding site).

735 To account for this, cluster-adjusted analyses used ‘de-clustered’ CpG sets that were constructed as follows:

- 736 i. Create CpG clusters formed from adjacent CpGs where each CpG is within 5kbp of the nearest
737 neighbouring CpG (see Appendix 1-Fig. 5 for a justification of this threshold);
- 738 ii. Construct de-clustered test set by randomly sampling a single CpG from each cluster; non-
739 clustered ‘singleton’ CpGs are always selected.

740 In the case of SoC-CpGs, the set of 259 non-clustered CpGs were reduced to 161 CpGs after de-clustering.

741 *Chromatin state and histone (H3K) mark analysis*

742 Data on chromatin states predicted by the ChromHMM 15-state model⁸⁵ in H1 ESCs, and 3 fetal tissues
743 (brain, muscle and small intestine) derived from 3 different germ layers, generated by the Roadmap
744 Epigenomics Consortium²⁹ was downloaded from the Washington University Roadmap Epigenomics

745 repository. 15-state model predictions were collapsed to 8 states for visualisation purposes, with sub-
746 classifications described in the Fig. 5 caption.

747 Overlaps with H3K marks for each of the above tissues were assessed using the *annotatr* (v1.10.0) package
748 in R to interrogate the same Roadmap Epigenomics ChIP-seq data used by ChromHMM for chromatin state
749 prediction.

750 *Additional modelling of seasonal variation in blood cell composition*

751 Cell count estimates using the Houseman method⁸⁶ were obtained using the *estimateCellCounts()* from *minfi*
752 (v1.30.0) in R. Seasonal variation in blood cell composition was then modelled by Fourier regression with
753 one pair of Fourier terms and sex (ENID+EMPHASIS) and age (EMPHASIS only) as adjustment covariates.
754 Fitted models indicated no marked seasonal differences within and between cohorts (Appendix 1-Fig. 17).

755 *Genetic association analyses*

756 Genotype data

757 mQTL and related SoC-association analyses were performed on all 284 individuals from the EMPHASIS (7-
758 9yr) cohort for which we had QC'd genotype data. 259 SoC-CpGs, plus sets of 259 matched and random
759 control CpGs were considered in this analysis. Subjects were genotyped using the Illumina Infinium Global
760 Screening Array-24 v1.0 Beadchip (Illumina, California, U.S.) following standard protocols. Array-derived
761 genotypes were pre-phased using SHAPEITv2 and imputation was performed using IMPUTEv2.3.2 on 1000
762 genomes phase 3 data. Further details are provided in Saffari et al⁷⁰. SNPs with a MAF $\leq 10\%$ were
763 excluded, along with those with an IMPUTE 'info' metric ≤ 0.9 , a stringent threshold to ensure maximum
764 confidence in imputation quality. Imputed SNPs were then pruned (using plink v1.90 -indep-pairwise with
765 window size 50, step size 5 and r^2 threshold of 0.8) to remove SNPs in strong LD. Finally, to minimise the
766 influence of low frequency homozygous variants in linear models, analysis was restricted to SNPs with 10 or
767 more homozygous variants, resulting in a final dataset comprising 2,609,310 SNPs.

768 Identification of mQTLs and 'GxE' SNPs

769 mQTL analysis was performed using the *GEM* package (v1.10.0) from R Bioconductor⁸⁷.

770 SNP effects on methylation were modelled as follows:

771 $M_j \sim \text{covs} + G$

772 Where M_j is the methylation M-value for CpG j , G is the SNP genotype coded as allelic dosage (0,1,2) and
773 covs correspond to the adjustment covariates used in the main EMPHASIS analysis (PCs 1-6, child age, sex
774 and intervention status).

775 GxE (SoC) effects were modelled as follows:

776 $M_j \sim \text{covs} + \sin\theta + G + G \times \sin\theta$

777 when $\sin\theta$ is the most significant Fourier term in the main SoC analysis,

778 OR

779 $M_j \sim \text{covs} + \cos\theta + G + G \times \cos\theta$

780 when $\cos\theta$ is the most significant Fourier term in the main SoC analysis.

781 Here, G and covs are as described above.

782 Since most mQTL effects are known to act in *cis*³², in order to maximise power a two-step procedure was
783 used to identify significant mQTLs:

784 i) *Identification of cis-mQTLs* passing FDR<5% considering the reduced set of SNPs within 1Mbp
785 of a CpG in the set to be analysed (SoC-CpGs, matched or random controls);

786 ii) *Identification of trans-mQTLs* passing FDR<5%, considering the full set of 2.6M SNPs.

787 Calculation of methylation variance explained by mQTLs

788 For each CpG j , total methylation variance explained was calculated for the following model:

789 $M_j \sim \text{covs} + \text{mQTL}_{j1} + \text{mQTL}_{j2} + \dots \text{mQTL}_{jn}$

where covs are as defined above, and $mQTL_{jn}$ is the genotype (coded 0,1,2) of the n th mQTL mapping to CpG j . Methylation variance explained was calculated from the model adjusted R^2 value, $adjR^2_{mQTL}$, to account for different levels of model complexity due to the differing number of mQTL identified for each CpG.

A final estimate of mQTL methylation variance explained was obtained by subtracting variance explained by the covariate-only model:

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

where $adjR^2_{cov}$ is the adjusted R^2 for the covariate-only model,

$$M_j \sim covs.$$

SoC association analysis

Potential confounding of SoC-DNA signals by SoC-associated genetic variants was assessed by analysing SoC associations with all 2,771 SoC-CpG-associated mQTLs identified in the mQTL analysis (see Table 4). SoC association analysis was performed under an allelic model using `-assoc` with *plink* v1.90. Significant mQTL-SoC associations were identified using $FDR < 5\%$ which assumes independence of all mQTL-SoC associations. We also considered a more liberal multiple testing correction threshold: $p_{Bonf} = 0.05/130$, assuming complete dependence of all *cis*-mQTL mapping to each of the 130 SoC-CpG with an associated mQTL.

Genetic ancestry sensitivity analysis

An investigation of population structure in the EMPHASIS cohort was conducted by first performing a principal component analysis (PCA) using `-pca` in *plink* v1.90. The PCA was performed on non-imputed data, with LD pruning using `-indep-pairwise 50 5 0.2` in *plink* v1.90, corresponding to an r^2 threshold of 0.2⁸⁸. Evidence for population structure was then obtained by plotting the first four principal components (Appendix 1-Fig. 12A). Confirmation of a likely link to Gambian ethnic ancestry in this largely ethnically Mandinka region of Gambia followed our observation that a distinct cluster (Appendix 1-Fig. 12B) was primarily made up of individuals from a single, predominantly ethnic Fula village.

A sensitivity analysis to check the effect of accounting for ancestry differences was performed by repeating the main analysis with the following additional covariates:

1. EMPHASIS (7-9yr) cohort

We adjusted for genetic ancestry directly using the first two genetic principal components identified in the genetic PCA.

2. ENID (2yr) cohort

Since no genetic data is available for this cohort, and since individuals from this cohort are drawn from the same villages as the EMPHASIS cohort on which we did the genetic PCA, we reasoned that village of origin is a useful proxy for genetic ancestry in our population. We therefore included an additional covariate dichotomised according to whether or not an individual was one of the 9 who came from the predominantly Fula village identified as genetic outlier in the EMPHASIS PCA analysis.

Sensitivity analysis to investigate confounding by infant sex

Our finding that multiple SoC-CpGs were associated with infant sex in previous EWAS prompted us to perform a sensitivity analysis checking for the possibility of a residual confounding effect due to sex. To do this we regressed out the effect of infant sex at each CpG in the M-value methylation matrix, prior to the main regression analysis. We then re-ran the Fourier regression analysis with and without an additional adjustment for infant sex in Fourier regression models. As expected, given that we adjusted for infant sex in the main analysis, this produced near identical results, suggesting that the main analysis was not confounded by infant sex.

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.

839 Signed informed consent for both studies was obtained from parents, and verbal assent was additionally
840 obtained from the older children who participated in the EMPHASIS study.

841

842 **References**

843

- 844 1. Smith, Z. D. & Meissner, A. DNA methylation: roles in mammalian development. *Nature*
845 *Reviews Genetics* **14**, 204–220 (2013).
- 846 2. Jeltsch, A. Molecular Enzymology of Mammalian DNA Methyltransferases. in *DNA*
847 *Methylation: Basic Mechanisms* 203–225 (Springer-Verlag). doi:10.1007/3-540-31390-7_7.
- 848 3. Feldmann, A. *et al.* Transcription Factor Occupancy Can Mediate Active Turnover of DNA
849 Methylation at Regulatory Regions. *PLoS Genetics* **9**, (2013).
- 850 4. Guo, H. *et al.* The DNA methylation landscape of human early embryos. *Nature* **511**, 606–610
851 (2014).
- 852 5. Fleming, T. P. *et al.* Origins of lifetime health around the time of conception: causes and
853 consequences. *The Lancet* **391**, 1842–1852 (2018).
- 854 6. Gonseth, S. *et al.* Periconceptional folate consumption is associated with neonatal DNA
855 methylation modifications in neural crest regulatory and cancer development genes. *Epigenetics*
856 **10**, 1166–76 (2015).
- 857 7. Tobi, E. W. *et al.* Early gestation as the critical time-window for changes in the prenatal
858 environment to affect the adult human blood methylome. *International Journal of*
859 *Epidemiology* **44**, 1211–1223 (2015).
- 860 8. Finer, S. *et al.* Is famine exposure during developmental life in rural Bangladesh associated
861 with a metabolic and epigenetic signature in young adulthood? A historical cohort study. *BMJ*
862 *Open* **6**, e011768 (2016).
- 863 9. Birney, E., Smith, G. D. & Grealley, J. M. Epigenome-wide Association Studies and the
864 Interpretation of Disease -Omics. *PLOS Genetics* **12**, e1006105 (2016).
- 865 10. Moore, S. E. *et al.* Prenatal or early postnatal events predict infectious deaths in young
866 adulthood in rural Africa. *International journal of epidemiology* **28**, 1088–95 (1999).

- 867 11. Dominguez-Salas, P. *et al.* DNA methylation potential: Dietary intake and blood concentrations
868 of one-carbon metabolites and cofactors in rural African women. *American Journal of Clinical*
869 *Nutrition* **97**, 1217–1227 (2013).
- 870 12. Moore, S. E. *et al.* A randomized trial to investigate the effects of pre-natal and infant
871 nutritional supplementation on infant immune development in rural Gambia: the ENID trial:
872 Early Nutrition and Immune Development. *BMC pregnancy and childbirth* **12**, 107 (2012).
- 873 13. Chandak, G. R. *et al.* Protocol for the EMPHASIS study; epigenetic mechanisms linking
874 maternal pre-conceptional nutrition and children's health in India and Sub-Saharan Africa.
875 *BMC Nutrition* **3**, 81 (2017).
- 876 14. Waterland, R. A. *et al.* Season of conception in rural gambia affects DNA methylation at
877 putative human metastable epialleles. *PLoS genetics* **6**, e1001252 (2010).
- 878 15. Dominguez-Salas, P. *et al.* Maternal nutrition at conception modulates DNA methylation of
879 human metastable epialleles. *Nature Communications* **5**, 1–7 (2014).
- 880 16. Silver, M. *et al.* Independent genomewide screens identify the tumor suppressor VTRNA2-1 as
881 a human epiallele responsive to periconceptional environment. *Genome Biology* **16**, 118 (2015).
- 882 17. Kessler, N. J., Waterland, R. A., Prentice, A. M. & Silver, M. J. Establishment of
883 environmentally sensitive DNA methylation states in the very early human embryo. *Science*
884 *Advances* **4**, eaat2624 (2018).
- 885 18. Van Baak, T. E. *et al.* Epigenetic supersimilarity of monozygotic twin pairs. *Genome Biology*
886 **19**, 2 (2018).
- 887 19. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in
888 mammals. *Trends in genetics : TIG* **18**, 348–51 (2002).
- 889 20. Anderson, O. S., Sant, K. E. & Dolinoy, D. C. Nutrition and epigenetics: an interplay of dietary
890 methyl donors, one-carbon metabolism and DNA methylation. *The Journal of Nutritional*
891 *Biochemistry* **23**, 853–859 (2012).
- 892 21. Gunasekara, C. J. & Waterland, R. A. A new era for epigenetic epidemiology. *Epigenomics* **11**,
893 1647–1649 (2019).

- 894 22. Rayco-Solon, P., Fulford, A. & Prentice AM. Differential effects of seasonality on preterm birth
895 and intrauterine growth. *American Journal of Clinical Nutrition* **81**, 134–139 (2005).
- 896 23. Sugden, K. *et al.* Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA
897 Methylation Measurement. *Patterns* **1**, 100014 (2020).
- 898 24. Okae, H. *et al.* Genome-Wide Analysis of DNA Methylation Dynamics during Early Human
899 Development. *PLoS Genetics* **10**, e1004868 (2014).
- 900 25. Monk, D., Mackay, D. J. G., Eggermann, T., Maher, E. R. & Riccio, A. Genomic imprinting
901 disorders: lessons on how genome, epigenome and environment interact. *Nature Reviews*
902 *Genetics* **20**, 235–248 (2019).
- 903 26. James, P. *et al.* Candidate genes linking maternal nutrient exposure to offspring health via DNA
904 methylation: a review of existing evidence in humans with specific focus on one-carbon
905 metabolism. *International Journal of Epidemiology* 1–28 (2018) doi:10.1093/ije/dyy153.
- 906 27. Zink, F. *et al.* Insights into imprinting from parent-of-origin phased methylomes and
907 transcriptomes. *Nature Genetics* **50**, 1542–1552 (2018).
- 908 28. Sanchez-Delgado, M. *et al.* Human Oocyte-Derived Methylation Differences Persist in the
909 Placenta Revealing Widespread Transient Imprinting. *PLOS Genetics* **12**, e1006427 (2016).
- 910 29. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–
911 330 (2015).
- 912 30. Waterland, R. A. & Jirtle, R. L. Transposable elements: targets for early nutritional effects on
913 epigenetic gene regulation. *Molecular and cellular biology* **23**, 5293–300 (2003).
- 914 31. Kazachenka, A. *et al.* Identification, Characterization, and Heritability of Murine Metastable
915 Epialleles: Implications for Non-genetic Inheritance. *Cell* 1–13 (2018)
916 doi:10.1016/j.cell.2018.09.043.
- 917 32. Gaunt, T. R. *et al.* Systematic identification of genetic influences on methylation across the
918 human life course. *Genome Biology* **17**, 61 (2016).

- 919 33. Hennig, B. J. *et al.* Cohort Profile: The Kiang West Longitudinal Population Study (KWLPs)—
920 a platform for integrated research and health care provision in rural Gambia. *International*
921 *Journal of Epidemiology* dyv206 (2015) doi:10.1093/ije/dyv206.
- 922 34. Ba, Y. *et al.* Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II
923 in maternal and cord blood. *Eur J Clin Nutr* **65**, 480–485 (2011).
- 924 35. Tobi, E. W. *et al.* Prenatal Famine and Genetic Variation Are Independently and Additively
925 Associated with DNA Methylation at Regulatory Loci within IGF2/H19. *PLoS One* **7**, (2012).
- 926 36. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association
927 studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012
928 (2019).
- 929 37. Do, C. *et al.* Genetic-epigenetic interactions in cis: A major focus in the post-GWAS era.
930 *Genome Biology* **18**, 1–22 (2017).
- 931 38. Comuzzie, A. G. *et al.* Novel Genetic Loci Identified for the Pathophysiology of Childhood
932 Obesity in the Hispanic Population. *PLOS ONE* **7**, e51954 (2012).
- 933 39. Evans, D. S. *et al.* Fine-mapping, novel loci identification, and SNP association transferability
934 in a genome-wide association study of QRS duration in African Americans. *Hum Mol Genet* **25**,
935 4350–4368 (2016).
- 936 40. van der Harst, P. *et al.* 52 Genetic Loci Influencing Myocardial Mass. *J Am Coll Cardiol* **68**,
937 1435–1448 (2016).
- 938 41. Kühnen, P. *et al.* Interindividual Variation in DNA Methylation at a Putative POMC Metastable
939 Epiallele Is Associated with Obesity. *Cell metabolism* **24**, 502–509 (2016).
- 940 42. Molaro, A. *et al.* Sperm methylation profiles reveal features of epigenetic inheritance and
941 evolution in primates. *Cell* **146**, 1029–1041 (2011).
- 942 43. Hammoud, S. S. *et al.* Distinctive chromatin in human sperm packages genes for embryo
943 development. *Nature* **460**, 473–478 (2009).
- 944 44. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers
945 in humans. *Nature* **470**, 279–285 (2011).

- 946 45. Vukic, M., Wu, H. & Daxinger, L. Making headway towards understanding how epigenetic
947 mechanisms contribute to early-life effects. *Philosophical Transactions of the Royal Society B:
948 Biological Sciences* **374**, 20180126 (2019).
- 949 46. Simpkin, A. J. *et al.* Longitudinal analysis of DNA methylation associated with birth weight
950 and gestational age. *Human Molecular Genetics* **24**, 3752–3763 (2015).
- 951 47. Teschendorff, A. E., West, J. & Beck, S. Age-associated epigenetic drift: implications, and a
952 case of epigenetic thrift? *Human Molecular Genetics* **22**, R7–R15 (2013).
- 953 48. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA
954 methylation data. *Nature Reviews Genetics* (2017) doi:10.1038/nrg.2017.86.
- 955 49. Bertozzi, T. M., Elmer, J. L., Macfarlan, T. S. & Ferguson-Smith, A. C. KRAB zinc finger
956 protein diversification drives mammalian interindividual methylation variability. *PNAS* (2020)
957 doi:10.1073/pnas.2017053117.
- 958 50. Carpenter, B. L. *et al.* Mother–child transmission of epigenetic information by tunable
959 polymorphic imprinting. *Proceedings of the National Academy of Sciences* 201815005 (2018)
960 doi:10.1073/pnas.1815005115.
- 961 51. Bogutz, A. B. *et al.* Evolution of imprinting via lineage-specific insertion of retroviral
962 promoters. *Nature Communications* **10**, 1–14 (2019).
- 963 52. Cavalli, G. & Heard, E. Advances in epigenetics link genetics to the environment and disease.
964 *Nature* **571**, 489–499 (2019).
- 965 53. Imbeault, M., Helleboid, P. Y. & Trono, D. KRAB zinc-finger proteins contribute to the
966 evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).
- 967 54. Shi, H. *et al.* ZFP57 regulation of transposable elements and gene expression within and beyond
968 imprinted domains. *Epigenetics & Chromatin* **12**, 49 (2019).
- 969 55. Monteagudo-Sánchez, A. *et al.* The role of ZFP57 and additional KRAB-zinc finger proteins in
970 the maintenance of human imprinted methylation and multi-locus imprinting disturbances.
971 *Nucleic Acids Res* doi:10.1093/nar/gkaa837.

- 972 56. Amarasekera, M. *et al.* Genome-wide DNA methylation profiling identifies a folate-sensitive
973 region of differential methylation upstream of ZFP57-imprinting regulator in humans. *FASEB*
974 *journal : official publication of the Federation of American Societies for Experimental Biology*
975 1–9 (2014) doi:10.1096/fj.13-249029.
- 976 57. Irwin, R. E. *et al.* A randomized controlled trial of folic acid intervention in pregnancy
977 highlights a putative methylation-regulated control element at ZFP57. *Clinical Epigenetics* **11**,
978 31 (2019).
- 979 58. Elliott, G. *et al.* Intermediate DNA methylation is a conserved signature of genome regulation.
980 *Nature Communications* **6**, 6363 (2015).
- 981 59. Tobi, E. W. *et al.* Selective Survival of Embryos Can Explain DNA Methylation Signatures of
982 Adverse Prenatal Environments. *Cell Reports* **25**, 2660–2667.e4 (2018).
- 983 60. Kang, L. *et al.* Aberrant allele-switch imprinting of a novel IGF1R intragenic antisense non-
984 coding RNA in breast cancers. *European Journal of Cancer* **51**, 260–270 (2015).
- 985 61. Sun, J. *et al.* A novel antisense long noncoding RNA within the IGF1R gene locus is imprinted
986 in hematopoietic malignancies. *Nucleic Acids Research* **42**, 9588–9601 (2014).
- 987 62. Boucher, J. *et al.* Insulin and insulin-like growth factor 1 receptors are required for normal
988 expression of imprinted genes. *Proceedings of the National Academy of Sciences* **111**, 14512–
989 14517 (2014).
- 990 63. Randhawa, R. & Cohen, P. The role of the insulin-like growth factor system in prenatal growth.
991 *Molecular Genetics and Metabolism* **86**, 84–90 (2005).
- 992 64. Aguirre, G. A., Ita, J. R., Garza, R. G. & Castilla-Cortazar, I. Insulin-like growth factor-1
993 deficiency and metabolic syndrome. *Journal of Translational Medicine* **14**, 1–23 (2016).
- 994 65. Larsson, O., Girnita, A. & Girnita, L. Role of insulin-like growth factor I receptor signalling in
995 cancer. *British Journal of Cancer* **92**, 2097–2101 (2005).
- 996 66. Tsai, P.-C. *et al.* DNA Methylation Changes in the IGF1R Gene in Birth Weight Discordant
997 Adult Monozygotic Twins. *Twin Research and Human Genetics* **18**, 635–646 (2015).

- 998 67. Singmann, P. *et al.* Characterization of whole-genome autosomal differences of DNA
999 methylation between men and women. *Epigenetics & Chromatin* **8**, 43 (2015).
- 1000 68. Shah, S. *et al.* Genetic and environmental exposures constrain epigenetic drift over the human
1001 life course. *Genome Research* **24**, 1725–1733 (2014).
- 1002 69. Suderman, M. *et al.* Sex-associated autosomal DNA methylation differences are wide-spread
1003 and stable throughout childhood. <http://biorxiv.org/lookup/doi/10.1101/118265> (2017)
1004 doi:10.1101/118265.
- 1005 70. Saffari, A. *et al.* Effect of maternal preconceptional and pregnancy micronutrient interventions
1006 on children’s DNA methylation: Findings from the EMPHASIS study. *The American Journal of*
1007 *Clinical Nutrition* nqaa193 (2020) doi:10.1093/ajcn/nqaa193.
- 1008 71. Kuehnen, P. *et al.* An Alu Element–Associated Hypermethylation Variant of the POMC Gene
1009 Is Associated with Childhood Obesity. *PLoS Genetics* **8**, e1002543 (2012).
- 1010 72. Candler, T. *et al.* DNA methylation at a nutritionally sensitive region of the PAX8 gene is
1011 associated with thyroid volume and function in Gambian children. *Science Advances* **7**,
1012 eabj1561 (2021).
- 1013 73. James, P. T. *et al.* Maternal One-Carbon Metabolism and Infant DNA Methylation between
1014 Contrasting Seasonal Environments: A Case Study from The Gambia. *Current Developments in*
1015 *Nutrition* **3**, nzy082 (2019).
- 1016 74. Hernandez-Vargas, H. *et al.* Exposure to aflatoxin B 1 in utero is associated with DNA
1017 methylation in white blood cells of infants in The Gambia. *International Journal of*
1018 *Epidemiology* **44**, 1238–1248 (2015).
- 1019 75. Logue, M. W. *et al.* The correlation of methylation levels measured using Illumina 450K and
1020 EPIC BeadChips in blood samples. *Epigenomics* **9**, 1363–1371 (2017).
- 1021 76. Xu, Z. & Taylor, J. A. Reliability of DNA methylation measures using Illumina methylation
1022 BeadChip. *Epigenetics* **16**, 495–502 (2021).
- 1023 77. Ek, W. E. *et al.* Genetic variants influencing phenotypic variance heterogeneity. *Human*
1024 *molecular genetics* **27**, 799–810 (2018).

- 1025 78. Owens, S. *et al.* Periconceptional multiple-micronutrient supplementation and placental
1026 function in rural Gambian women: A double-blind, randomized, placebo-controlled trial.
1027 *American Journal of Clinical Nutrition* **102**, 1450–1459 (2015).
- 1028 79. Min, J. L., Hemani, G., Davey Smith, G., Relton, C. & Suderman, M. Meffil: efficient
1029 normalization and analysis of very large DNA methylation datasets. *Bioinformatics (Oxford, England)* **34**, 3983–3989 (2018).
- 1031 80. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina
1032 Infinium HumanMethylation450 microarray. *Epigenetics : official journal of the DNA Methylation Society* **8**, 203–9 (2013).
- 1034 81. Fortin, J. P. *et al.* Functional normalization of 450k methylation array data improves replication
1035 in large cancer studies. *Genome Biology* **15**, 0–42 (2014).
- 1036 82. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels
1037 by microarray analysis. *BMC bioinformatics* **11**, 587 (2010).
- 1038 83. Zheng, Y. *et al.* Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Research* **45**, 8697–8711 (2017).
- 1040 84. Crider, K. S., Yang, T. P., Berry, R. J. & Bailey, L. B. Folate and DNA Methylation: A Review
1041 of Molecular Mechanisms and the Evidence for Folate's Role². *Adv Nutr* **3**, 21–38 (2012).
- 1042 85. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization.
1043 *Nature Methods* **9**, 215–216 (2012).
- 1044 86. Jaffe, A. E. & Irizarry, R. a. Accounting for cellular heterogeneity is critical in epigenome-wide
1045 association studies. *Genome biology* **15**, R31 (2014).
- 1046 87. Pan, H., Holbrook, J. D., Karnani, N. & Kwoh, C. K. Gene, Environment and Methylation
1047 (GEM): a tool suite to efficiently navigate large scale epigenome wide association studies and
1048 integrate genotype and interaction between genotype and environment. *BMC Bioinformatics* **17**,
1049 299 (2016).
- 1050 88. Weale, M. E. Quality control for genome-wide association studies. *Methods Mol. Biol.* **628**,
1051 341–372 (2010).

1052 89. Silver. *SoCFourier*. (Zenodo, 2021). doi:10.5281/ZENODO.5801480.

1053

1054

1055 **Acknowledgments**

1056 **Funding:** The Gambian ENID trial was jointly funded by the UK Medical Research Council
1057 (MRC) and the Department for International Development (DFID) under the MRC/DFID
1058 Concordat agreement (MRC Program MC-A760-5QX00). Methylation analysis of ENID samples
1059 was supported by the Bill & Melinda Gates Foundation (grant no: OPP1 066947). The Gambian
1060 EMPHASIS study is jointly funded by MRC, DFID and the Department of Biotechnology, Ministry
1061 of Science and Technology, India under the Newton Fund initiative (MRC grant no.:
1062 MR/N006208/1 and DBT grant no.: BT/IN/DBT-MRC/DFID/24/GRC/2015–16). Further support
1063 for this analysis was provided by MRC Grant MR/M01424X/1. We acknowledge the work of the
1064 full EMPHASIS Study Group (www.emphasisstudy.org) in acquiring this data.
1065

1066 **Disclaimer:** Where authors are identified as personnel of the International Agency for Research on
1067 Cancer / World Health Organization, the authors alone are responsible for the views expressed in
1068 this article and they do not necessarily represent the decisions, policy or views of the International
1069 Agency for Research on Cancer / World Health Organization.
1070

1071 **Competing interests:** Authors declare that they have no competing interests.
1072

1073 **Data and materials availability:**

1074 Data and code availability

1075 Illumina 450k methylation array data generated from Gambian 2 year olds from the ENID trial is
1076 deposited in GEO (GSE99863). Requests to access and analyse the other Gambian methylation
1077 datasets (ENID 5-7yr and EMPHASIS 7-9yr) should be submitted to the corresponding author in
1078 the first instance. An application would then need to be made to MRC Unit The Gambia's Scientific
1079 Coordinating Committee and the Joint MRC/Gambia Government Ethics Committee. Sources and
1080 locations of other publicly available data used in this analysis are described in Methods. Bespoke
1081 code used in the analysis is available at <https://zenodo.org/record/5801480>⁸⁹.
1082
1083
1084
1085

Figures captions

Figure 1. Study design. DNAm: DNA methylation; EPIC: Illumina Infinium MethylationEPIC BeadChip; Epig. Roadmap: Roadmap Epigenomics Consortium; ESC: embryonic stem cell; ESS: epigenetic supersimilarity; GSA: Global Screening Array; HM450: Illumina Infinium HumanMethylation450 BeadChip; IVF: in vitro fertilisation; MEs: metastable epialleles; MZ/DZ: monozygotic / dizygotic twins; PofO: parent of origin; RRBS: reduced representation bisulfite-seq; SIV: systemic interindividual variation; SoC: season of conception; WGBS: whole genome bisulfite-seq. See Tables 1 and 3 for further details of Gambian and public datasets used in this analysis.

Figure 2. Identification of Gambian season of conception associated CpGs: ENID (2yr) vs EMPHASIS (7-9yr) cross-cohort analysis. (A) Relationship between date of conception and date of sample collection for ENID (top) and EMPHASIS (bottom) cohorts. (B) Modelled seasonal change in methylation for 768 SoC-associated loci ($FDR < 5\%$) in the ENID cohort. 26 ME CpGs are marked in red. (C) Conception date of modelled methylation maximum in each cohort for 61 CpGs significantly associated with SoC in both cohorts (left) and 61 matched controls (right). (D) (Left) Date of modelled DNAm maximum vs seasonal amplitude in each cohort for 768 CpGs significantly associated with SoC in the ENID cohort. MEs are marked in red. Dashed line indicates SoC amplitude threshold used to identify SoC-CpGs. (Right) same CpGs as left, but in the older EMPHASIS cohort. Significant SoC association for this cohort are marked in a darker colour. (E) (Left) Seasonal amplitudes for SoC-associated CpGs that are (red) or are not (blue) MEs; along with amplitudes for 768 matched and random controls (light/dark grey respectively). (Right) as left but in the older EMPHASIS cohort. For EMPHASIS significant SoC associations are marked in a darker colour. Boxes represent the middle 50% of the data (interquartile range, IQR); the line inside the box is the median, and whiskers represent values lying within 1.5 times the IQR.

Figure 3. Identification of Gambian season of conception associated CpGs: ENID 2yr vs 5-7yr longitudinal analysis. (A) Relationship between conception date of modelled methylation maximum measured at 2yr and 5-7yr in the same $n=138$ individuals from the ENID cohort. $n=157$ SoC-CpGs with a significant SoC association ($FDR < 5\%$) at 5-7yr are plotted. (B) Change in seasonal amplitude between 2yr and 5-7yr for $n=259$ SoC-CpGs (left) and matched controls (right) with DNAm measured in the same $n=138$ individuals from the ENID cohort. There is evidence of SoC effect attenuation with age at SoC-CpGs, but not at matched controls (Wilcoxon signed rank test for difference in SoC amplitude $p=6.7 \times 10^{-12}$ and $p=1.0$ respectively).

Figure 4. Properties of SoC-CpGs. (A) Date of conception at modelled methylation maxima for 259 SoC-CpGs and 259 corresponding matched and random controls across all 3 analysed datasets. Green and yellow bands indicate the extent of the rainy and dry seasons respectively. (B) SoC-CpGs show increased intermediate methylation compared to array background (data from ENID 2yr and EMPHASIS cohorts combined). (C) Distribution of 259 SoC-CpGs (including MEs), 259 matched controls and array background with respect to CpG islands. Error bars are bootstrapped 95% CIs. N / S Shore / Shelf: North / South Shore / Shelf respectively (regions proximal to CpG Islands defined in Illumina manifest). (D) Distribution of pairwise Spearman correlations for CpG sets in the ENID (left) and EMPHASIS (right) cohorts. Boxplot elements as described in Fig. 2E.

Figure 5. Early stage embryo, gametic and parent-of-origin specific methylation. (A) Methylation distribution of SoC-CpGs, matched controls and array background in pre-gastrulation inner cell mass (ICM) and post-gastrulation embryonic liver, measured in RRBS embryo methylation data from Guo *et al*⁴. Data comprises 112,380 CpGs covered at $\geq 10x$ in ICM and/or embryonic liver that overlap array background, including 118 SoC-CpGs and 51 matched controls. (B) Methylation distribution of SoC-CpGs, matched controls and array background in sperm WGBS data from Okae *et al*²⁴. Data comprises 294,240 CpGs covered at $\geq 10x$ including 196 SoC-CpGs and 207 matched controls. (C) Sperm methylation and oocyte gDMR (oo-gDMR) status at 196 SoC-CpGs covered at $\geq 10x$ in Okae *et al* sperm WGBS data. Sperm hypomethylation is defined as methylation $\leq 25\%$. oo-gDMRs defined as sperm methylation $< 25\%$ and oocyte methylation $> 75\%$ in WGBS analysis by Sanchez-Delgado. *et al*²⁸. (D) Mean methylation at SoC-CpGs and array background measured in $n=233$ individuals in the ENID (2yr) cohort stratified by sperm and oocyte methylation status. Left: Methylation stratified according to sperm hypomethylation status ($n=175$ SoC-CpGs hypomethylated in sperm, $n=21$ not hypomethylated). Right: as left but with loci hypomethylated in sperm further stratified according to oo-gDMR status ($n=26$ SoC-CpGs hypermethylated in oocytes / oo-

1144 gDMR, n=149 not hypermethylated). Sperm / oo-gDMR status thresholds as for Fig. 4C. 8 PofOm CpGs
1145 (green triangles) are those identified in Zink *et al*²⁷.
1146

1147 **Figure 6. SoC-CpG overlap with predicted chromatin states.** Chromatin states predicted by
1148 ChromHMM⁸⁵ from chromatin marks in four cell lines and tissues generated by the Roadmap Epigenomics
1149 Consortium²⁹. Predicted states for all 259 SoC-CpGs are shown. Predictions from the ChromHMM 15-state
1150 model are collapsed to 8 states for clarity. TSS: active transcription start site/flanking active TSS/bivalent or
1151 poisedTSS; TSS/Enhancer: flanking bivalent TSS/enhancer; Enhancer: Enhancer/bivalent enhancer/Genic
1152 enhancer; Transcription: Strong/weak transcription/ Transcription at gene 5' and 3'; ZNF/repeat: zinc finger
1153 genes & repeats; Polycomb: Repressed/ weak repressed polycomb; Heterochromatin; Low signal: low signal
1154 in all marks states used as inputs to ChromHMM.
1155

1156 **Figure 7. Links between endogenous retroviruses (ERVs), ZFP57 binding sites, genetic**
1157 **variation and DNAm at SoC-associated loci.** (A) Proportion of SoC-CpGs, matched controls and
1158 array background CpGs proximal to ERV1 endogenous retroviral elements (top) and ZFP57
1159 binding sites (bottom), within the specified distance. CpG clustering effects are removed by
1160 sampling a single CpG from each cluster (see Methods). Error bars are bootstrapped 95% CIs. (B)
1161 Proportion of methylation variance explained by mQTL for matched controls, SoC-CpGs and
1162 random controls. Only CpGs with at least one significant mQTL are plotted (n=201, 130 and 50
1163 respectively; see Methods for further details). Boxplot elements as described in Fig. 2E.
1164

Tables

Table 1. Gambian seasonality-methylation analysis: cohort characteristics

ENID (2yr) ¹²	233	2yr	50.6	peripheral blood	Illumina Infinium HM450
ENID (5-7yr)	138	5-7yr	56.5		Illumina Infinium MethylationEPIC
EMPHASIS ¹³	289	7-9yr	54.3	peripheral blood	Illumina Infinium MethylationEPIC

Note: Individuals with ENID longitudinal (5-7yr) methylation data are a subset of those with methylation at 2yr. There is no overlap between individuals included in the ENID and EMPHASIS cohorts.

1175 **Table 2. CpG sets considered in this analysis.**

1176

Array background	391,814	Intersection of CpGs on Illumina HM450 (ENID 2yr) and EPIC (EMPHASIS) cohort arrays, post QC
SoC-CpGs	259	SoC-associated CpGs with SoC effect size (SoC methylation amplitude) > 4% in the ENID 2yr dataset
Matched controls	259	CpGs with similar methylation distributions to SoC-CpGs in the ENID 2yr dataset*
Random controls	259	Random sample from array background

1177

1178 *Matching methylation distributions determined by Kolmogorov-Smirnov tests (see Appendix 1-Fig. 16). QC:
 1179 quality control; LRT: likelihood ratio test. See Methods for further details.

1180

1181

1182 **Table 3. External datasets considered in this analysis.**

1183

Putative metastable epialleles (MEs)	1881 ME/SIV/ESS CpGs overlapping array background identified in multi-tissue and MZ/DZ screens in Van Baak <i>et al</i> ¹⁸ and Kessler <i>et al</i> ¹⁷ .
Parent-of-origin specific methylation (PofOm)	699 Parent-of-origin specific methylation loci identified in peripheral blood in Zink <i>et al</i> ²⁷ overlapping array background.
Embryo DNAm data	RRBS data for inner cell mass and embryonic liver (<10 weeks gestation) from Guo <i>et al</i> ⁴ .
Sperm DNAm data	WGBS data from Okae <i>et al</i> ²⁴ .
Germline DMRs (gDMRs)	Regions differentially methylated in sperm and oocytes identified in WGBS data by Sanchez-Delgado <i>et al</i> ²⁸ .
Transposons (ERVs)	ERVs determined by RepeatMasker were downloaded from the UCSC h19 annotations repository.
Transcription factor ChIP-seq	ZFP57, TRIM28 and CTCF transcription factor binding sites identified from ChIP-seq in human embryonic kidney and hESCs are described in Kessler <i>et al</i> ¹⁷ .
Chromatin state predictions and histone 3 marks	Chromatin state predictions for H1 ESCs, fetal brain, fetal muscle and fetal small intestine generated using ChromHMM ⁸⁵ from Roadmap Epigenomic Consortium data ²⁹ . Histone mark data are from the same source.

1184

1185

1186

1187

1188

1189

1190

1191

1192

ME: metastable epiallele; SIV: systemic interindividual variation; ESS: epigenetic supersimilarity; MZ/DZ: monozygotic/dizygotic twins; PofOm: parent-of-origin methylation; RRBS: reduced-representation bisulfite-seq; DMR: differentially methylated region; ERV: endogenous retrovirus; ESCs: embryonic stem cells. See Methods for further details.

1193 **Table 4. Methylation quantitative trait loci (mQTL) associated with SoC-CpGs and controls**
1194

SoC-CpGs	130 (50%)	2771 (2549/222)	6 (2-30)	0.09 (0.08-0.15)
Matched controls	201 (78%)	7886 (7417/469)	15 (4-50)	0.09 (0.06-0.21)
Random controls	50 (19%)	1512 (1476/36)	7 (2-35)	0.1 (0.08-0.18)

1195
1196 ¹ delta adjusted R² (see Methods); IQR: inter-quartile range.
1197

1198 **Supplementary Files**

1199

1200 Appendix 1 contains Supplementary Figures 1-17

1201

1202 Supplementary File1 contains Supplementary Tables 1a-1s

1203 1a ENID SoC-associated CpGs (no amplitude filter applied)

1204 1b Seasonal amplitude tests

1205 1c Inter-cohort change in SoC amplitude at ENID SoC-associated CpGs

1206 1d SoC-CpGs (ENID SoC-associated CpGs with SoC methylation amplitude $\geq 4\%$)

1207 1e SoC-CpG clusters and singletons

1208 1f SoC-CpG cluster sizes

1209 1g SoC-CpG enrichment for MEs and overlap with ME sub-classes

1210 1h SoC-CpG enrichment for gDMRs and parent-of-origin specific methylation

1211 1i SoC-CpGs overlapping maternally methylated germline DMRs

1212 1j SoC-CpGs mQTL and their association with SoC

1213 1k Candidate genes previously associated with periconception / gestational exposures

1214 1l Look-up of SoC-CpGs in the EWAS Catalog

1215 1m EWAS Catalog data grouped by trait

1216 1n SoC-CpG associations with selected variables measured in the ENID 2yr dataset

1217 1o Look-up of SoC-CpGs in the GWAS Catalog

1218 1p ENID (2yr): Association test p-values to detect potential residual confounding

1219 1q EMPHASIS (7-9yr): Association test p-values to detect potential residual confounding

1220 1r ENID 5-7yr: Association test p-values to detect potential residual confounding

1221 1s ENID (2yr): Cell count, batch, village sensitivity analyses

Gambian season of conception analysis

Methylation
'hotspots' sensitive
to early embryo
environment

Tools to probe peri-conceptual mechanisms

A **C**

H1 ESC

fetal brain

fetal muscle

fetal small intestine

Appendix 1-Figure 1. Cross-cohort correlation of mean methylation values at 768 SoC-associated loci. Data comprises n=233 and n=289 individuals from the ENID (2yr) and EMPHASIS cohorts respectively. 768 SoC-associated CpGs (FDR<5%) identified in the ENID (2yr) cohort.

Appendix 1-Figure 2. Date of conception vs date of collection for samples analysed in the ENID longitudinal (5-7yr) dataset. Data comprises samples from n=138 ENID participants with methylation data at both 2yr and 5-7yr.

Appendix 1-Figure 3. Date of conception at modelled methylation minima. Conception dates at methylation minima for loci plotted in Fig. 4A. Note that since seasonality is modelled by a single pair of Fourier terms, maxima and minima are 6 months apart.

Appendix 1-Figure 4. Genomic locations (hg19) of 259 SoC-CpGs.
Data from Appendix 1-Table 4.

Appendix 1-Figure 5. Relationship between the inter-CpG distance used to define SoC-CpG clusters and the number of identified clusters and singletons (CpGs not falling within a cluster).

Appendix 1-Figure 6. Comparison of SoC effects at SoC-CpGs within and outside of SoC-CpG clusters. Top: Conception date at modelled methylation maxima for ENID 2yr (x-axis) and EMPHASIS (y-axis) cohorts, according to whether locus falls within (n=154; left-red), or outside (n=105; right-blue) of a SoC-CpG cluster. **Bottom:** Seasonal effect size (amplitude) for ENID 2yr (left) and EMPHASIS (cohorts), according to whether locus falls within (red), or outside (blue) of a SoC-CpG cluster. Numbers are p-values for two-sided Wilcoxon Rank Sum tests under the null hypothesis of no difference between the two distributions.

16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52

53
54
55
56
57
58

Appendix 1-Figure 7. Seasonal variation at the *IGF1R* locus in ENID (2yr) and EMPHASIS cohorts. Unadjusted methylation Beta values are plotted. Boxes represent IQRs for season-specific DNAm at each locus. Note that for visualisation purposes Gambian seasons are dichotomised (Dry: Jan-Jun; Rainy: Jul-Dec), whereas seasonality is modelled as a continuous cyclical variable in Fourier regression models used in the main analysis.

Appendix 1-Figure 8. Cluster-adjusted pairwise inter-CpG methylation correlations for CpG sets in ENID 2yr (left) and EMPHASIS (right) cohorts. This is the same as Fig. 4D, but with a single CpG randomly sampled from each CpG cluster so that CpGs in each set are a minimum distance of 5000bp apart. This shows that pairwise correlation distributions are not disproportionately driven by large numbers of pairwise correlations between highly-correlated neighbours.

289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323

Appendix 1-Figure 9. Mean methylation at SoC-CpGs and array background measured in n=289 individuals in the EMPHASIS (7-9yr) cohort stratified by sperm and oocyte methylation status. (Left): Methylation stratified according to sperm methylation status reported in Okae *et al.*²³. Sperm hypomethylation is defined as methylation $\leq 25\%$. **(Right)** as left but with loci hypomethylated in sperm further stratified according to oocyte gDMR (oo-gDMR) status reported in Sanchez-Delgado *et al.*²⁷. PofOm CpGs are those identified in Zink *et al.*²⁶ oo-gDMR have oocyte methylation $>75\%$.

Broad peaks

(259 SoC-CpGs with annotations)

Narrow peaks

(97 SoC-CpGs with annotations)

Appendix 1-Figure 10. Histone marks overlapping SoC-CpGs in H1 embryonic stem cells. Histone marks or combinations thereof are ordered by abundance. H3 marks were generated by the Roadmap Epigenomics Consortium²⁸ and downloaded using the *annotatr* (v1.10.0) package in R. Broad peaks overlapping all 259 SoC-CpGs (**top**) and (more sharply defined) narrow peaks overlapping 97 SoC-CpGs (**bottom**) correspond to different thresholds used by the MACS peak-calling algorithm. See Kundaje et al²⁸ for further details.

Appendix 1-Figure 11. Links between ERVK endogenous retroviral elements, ZFP57 and TRIM28 binding sites, and DNAm at SoC-CpGs. (A) Proportion of SoC-CpGs, matched controls and array background CpGs proximal to ERVK endogenous retroviral elements (top), and ZFP57 and TRIM28 binding sites (bottom), within the specified distance. CpG clustering effects are removed by sampling a single CpG from each cluster (see Methods). Error bars are bootstrapped 95% CIs.

Appendix 1-Figure 12. Population structure in the EMPHASIS cohort.

A: 1st and 2nd (left) and 2nd and 3rd (right) principal components (PCs) from a PCA of genome-wide genetic data from n=294 individuals from the EMPHASIS cohort. Figures in brackets give % variance explained by each PC. Colours correspond to the 24 villages from the Kiang West region which are covered in this cohort. A majority of individuals are of Mandinka origin and villages that are predominantly Mandinka are tightly clustered in the PCA plot. The distinct cluster in the top left of the first plot corresponds to individuals from a single village which is predominantly Fula. **B:** Same as **A** left, but with individuals from the 'outlier village' plotted as a distinct colour. This illustrates that all 16 individuals from this village are distinct from the main cluster.

Appendix 1-Figure 13. Season of conception analysis adjusted for ethnicity. This figure replicates Fig. 4A in the main paper for the ENID 2yr and EMPHASIS cohorts, but with additional adjustment for ethnicity. Here we adjust for ethnicity in the EMPHASIS cohort using the first two genetic principle components as covariates in the Fourier regression models (Appendix 1-Fig. 12A). For the ENID cohort where we do not have genetic data, we adjust for ethnicity using an additional covariate dichotomised according to whether an individual was one of the 9 who came from the predominantly Fula ‘outlier village’ identified from the PCA using genetic data in the EMPHASIS cohort (Appendix 1-Fig. 12B).

Appendix 1-Figure 14. UCSC genome browser plot of the *IGF1R* SoC-CpG cluster. 7 SoC-CpGs on the Illumina 450k array are highlighted. This region falls within intron 2 and bears the hallmarks of being a promoter and/or active or poised enhancer in multiple cell lines. The key lists colour coding for predicted regulatory regions using ChromHMM.

Appendix 1-Figure 15. Seasonal differences in methylation at 30,573 LINE1 and 30,840 Alu elements, and 391,814 array background CpGs. Methylation values at each locus are centred to have mean zero to enable comparison across loci. LINE1 and Alu methylation values are predicted using REMP (v1.16.0) (see Methods). Season of conception is dichotomised to dry: Jan-June; rainy: July-December.

586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635

Appendix 1-Figure 16. Selection of matched controls. Matched controls are selected from array background using Kolmogorov-Smirnov (KS) tests to identify CpGs with similar methylation distributions to SoC-CpGs (see Methods). **Top:** Methylation mean distribution at 259 SoC-CpGs (left) vs 259 matched controls (right). **Bottom:** Methylation distribution of 10 SoC-CpGs selected at random (thick red lines) and their corresponding KS-matched controls (thin black lines).

36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74

Appendix 1-Figure 17. Seasonal variation in estimated cell composition. Cell proportions for all samples from each cohort are estimated using the estimateCellCounts function from the *minfi* package in R. Seasonal variation (red curves) for each cell type is determined from Fourier regression models adjusted for sex (ENID) or age and sex (EMPHASIS). See Methods for further details.