bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DNA methylation covariation in human whole blood and sperm: implications for studies of intergenerational epigenetic effects

Fredrika Åsenius[1], Tyler J. Gorrie-Stone[2], Ama Brew[3], Yasmin Panchbaya[4], Elizabeth Williamson[5], Leonard C. Schalkwyk[2], Vardhman K. Rakyan[3], Michelle L. Holland[6], Sarah J. Marzi[7,8,#,+], David J. Williams[1,#]

[1] UCL EGA Institute for Women's Health, University College London, London, UK

[2] School of Biological Sciences, University of Essex, Colchester, UK

[3] The Blizard Institute, Queen Mary University of London, London, UK

[4] UCL Genomics, Great Ormond Street Institute of Child Health, London, UK

[5] Fertility & reproductive medicine laboratory, University College Hospital, London, UK

[6] Department of Medical and Molecular Genetics, School of Basic and Medical Biosciences, King's College London, London, UK

[7] UK Dementia Research Institute, Imperial College London, London, UK

[8] Department of Brain Sciences, Imperial College London, London, UK

[#] Joint senior authors: Sarah J. Marzi, David J. Williams

[+] **Correspondence should be addressed to Sarah J. Marzi (s.marzi@imperial.ac.uk)**

Abstract

Background

Epidemiological studies suggest that paternal obesity may increase the risk of fathering small for gestational age offspring. Studies in non-human mammals suggest that such associations could be mediated by DNA methylation changes in spermatozoa that influence offspring development in utero. Human obesity is associated with differential DNA methylation in peripheral blood. It is unclear, however, whether this differential DNA methylation is reflected in spermatozoa. We profiled genome-wide DNA methylation using the Illumina MethylationEPIC array in matched human blood and sperm from lean (discovery n=47; replication n=21) and obese (n=22) males to analyse tissue covariation of DNA methylation, and identify whether this covariation is influenced by obesity.

Results

DNA methylation signatures of human blood and spermatozoa are highly discordant, and methylation levels are correlated at only a minority of CpG sites (~1%). While at the majority of these sites, DNA methylation appears to be influenced by genetic variation, obesity-associated DNA methylation in blood was not generally reflected in spermatozoa, and obesity did not influence covariation patterns. However, one cross-tissue obesity-specific hypermethylated site (cg19357369; chr4:2429884; P=8.95 $\times$ 10^{-8}; beta=0.02) was identified, warranting replication and further investigation. When compared to a wide range of human somatic tissue samples (n=5,917), spermatozoa displayed differential DNA methylation in pathways enriched in transcriptional regulation.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Conclusions

Human sperm displays a unique DNA methylation profile that is highly discordant to, and practically uncorrelated with, that of matched peripheral blood. Obesity only nominally influences sperm DNA methylation, making it an unlikely mediator of intergenerational effects of metabolic traits.

Keywords

DNA methylation, sperm, intergenerational inheritance, obesity, epigenetics

Background

Multiple large-scale epigenome-wide association studies in humans have shown that environmental and acquired phenotypes, including smoking, ageing and obesity, are associated with altered DNA methylation in peripheral blood [1-4]. Whether such phenotypes also have the potential to induce epigenetic changes in gametes has generated considerable interest in recent years. Studies in non-human mammals suggest that the spermatozoal DNA methylome can be influenced by factors such as dietary alterations, toxicants and even psychological stress [5-10], although the majority of these results have yet to be replicated independently. A small number of studies also suggest that acquired traits in male mice induce epigenetic changes in sperm, which in turn influence the physiology of offspring [7, 11, 12].

There is little evidence for such inter- and transgenerational effects of acquired phenotypes via epigenetic inheritance in humans. This is partly due to the fact that human sperm is rarely analysed outside of a reproductive medicine setting and is less accessible than, for example, peripheral blood. Further, it is ethically and practically impossible to perform a study of transgenerational effects in humans in which all potential external and lifestyle-related confounders are removed, and inter-individual genetic variation is generally not controllable. In addition, one needs to account for the two-stage process of epigenetic reprogramming of primordial germ cells and preimplantation embryos that occurs between generations [13]. Lastly, epigenetic signatures are highly tissue- and developmental stage specific [14, 15], making findings from studies using whole blood as a surrogate tissue for spermatozoa difficult to interpret [16].

Despite these caveats, epidemiological evidence suggests that factors such as advanced paternal age, obesity, diabetes and smoking have the potential to negatively impact the development and physiology of a man's offspring [17-19], presumably via alterations to his spermatozoa (**Figure 1A**). An improved understanding of whether and how acquired paternal traits can influence offspring physiology has important implications, both scientifically and in terms of public health policy. This is particularly pertinent for modifiable traits such as obesity, where timely intervention could reduce any potential negative intergenerational effects.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

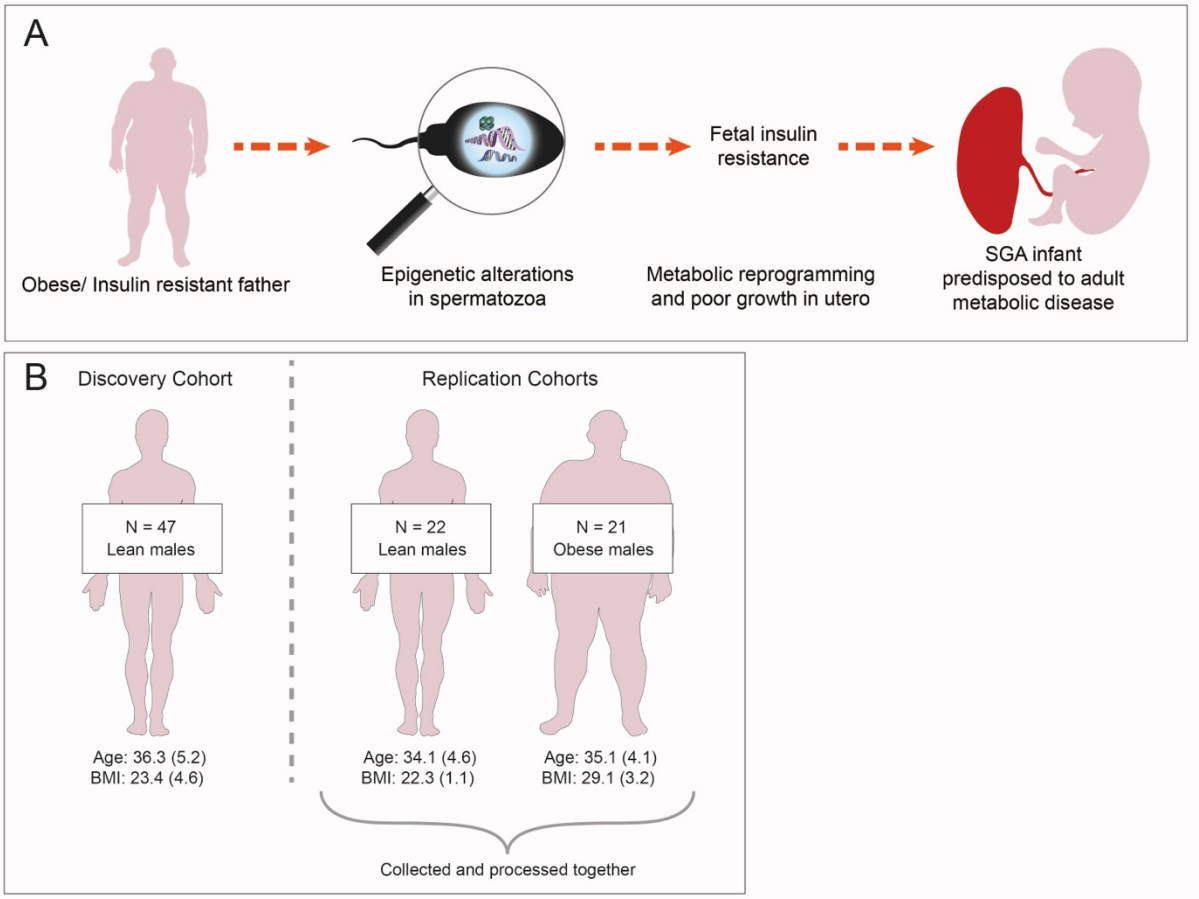


Figure 1. Intergenerational epigenetic inheritance via spermatozoa and overview of study cohorts

A) Mechanism for how acquired paternal phenotypes could alter offspring physiology via epigenetic alterations to a man's spermatozoa. Epidemiological studies suggest that some acquired paternal traits, including obesity and insulin resistance, are associated with an increased risk of fathering small for gestational age (SGA) offspring [18, 19, 58]. Studies in non-human mammals suggest that such associations could be mediated by DNA methylation alterations in spermatozoa that induce metabolic reprogramming in the developing foetus [12].

B) Overview of study cohorts. The discovery cohort included 47 lean males (BMI 19-25 kg/m^2) and the replication cohorts included 22 lean males (BMI 19-25 kg/m^2) and 21 overweight/obese males (BMI >26 kg/m^2; 'the obesity cohort'). Age (years) and BMI (kg/m^2) are expressed as mean (SD).

SGA: small for gestational age. SD: standard deviation.

It will be a long time before studies of DNA methylation in human spermatozoa reach a comparable magnitude to those currently available on peripheral blood. Therefore, it is of interest to identify CpG sites where DNA methylation levels covary between the two tissues, that is, sites at which blood methylation is predictive of sperm methylation, even if the absolute level of methylation is different. The extent to which these sites overlap with those identified in blood as associated with environmental stimuli or acquired phenotypes will provide new insight into whether the sperm methylome may be similarly responsive. At such CpG sites, using blood DNA methylation as a proxy for inferring DNA methylation in spermatozoa might be justified. To our knowledge, the largest study that analysed genome-wide DNA methylation in an unbiased manner in matched samples of blood and sperm to date included a total of eight participants [20].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In this study, we analysed genome-wide DNA methylation using the Infinium MethylationEPIC array in matched samples of human blood and sperm from lean (n = 68) and overweight/obese (n = 22; 'the obesity cohort') healthy males of proven fertility. We interrogated the extent to which obesity-associated DNA methylation in blood is reflected in spermatozoa from obese males and identified obesity associated CpG-sites in sperm and blood. Spermatozoal DNA methylation data was further compared to that of nearly 6,000 somatic tissue samples available on the Gene Expression Omnibus data repository [21], allowing us to identify sperm-specific DNA methylation signatures. Together, our analyses interrogate the plausibility of spermatozoal DNA methylation as a mechanism for intergenerational effects of paternal obesity and whether whole blood can be used as a surrogate tissue for analyses of DNA methylation when sperm is unavailable. Further, they provide a unique insight into how spermatozoal DNA methylation compares to DNA methylation in a wide range of human somatic tissues.

Results

General characterisation of the sperm DNA methylome

We used the Illumina MethylationEPIC array to quantify DNA methylation at > 850,000 CpG sites across the human genome in matched samples of whole blood and sperm from a discovery cohort of 47 lean, healthy males of proven fertility. Following pre-processing, normalization and stringent quality control (see **Materials and Methods**), a total of 704,356 probes were retained for further analyses. Raw and pre-processed DNA methylation data is available for download from the Gene Expression Omnibus (GEO) at accession number GSE149318. To characterize spermatozoal DNA methylation across genomic regions, levels of DNA methylation were divided into three categories; 'low', 'intermediate' and 'high', corresponding to median beta values < 0.2, 0.2-0.8 and > 0.8 across individuals respectively (**Figure 2**). As observed in other tissues and cell types, CpG islands and shores generally show low DNA methylation in sperm. Conversely, sites mapping to the open sea were characterized by overall higher DNA methylation (**Figure 2A**, **Table S1**). Gene bodies in spermatozoa displayed overall high levels of DNA methylation, whilst sparser DNA methylation was seen around transcription start sites (TSS) and 5' untranslated regions (UTRs), as well as the first exons (**Figure 2B**, **Table S2**).

In line with previous reports, we confirmed that the DNA methylation age estimator developed by Horvath [4] worked well in whole blood ($r = 0.74$, $P = 2.55 \times 10^{-9}$, Pearson's product moment correlation), but not in sperm ($r = 0.26$, $P = 0.07$, **Figure S1A**). This is likely because the Horvath DNA methylation was developed using only 45 samples of semen in a total of 7,844 samples (0.6%) of different tissue samples, including 4,180 blood-derived samples (53%) [4]. However, age could more accurately be predicted using the model recently developed by Jenkins and colleagues [22], which was specifically trained on sperm samples ($r = 0.68$, $P = 1.78 \times 10^{-7}$, **Figure S1B**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

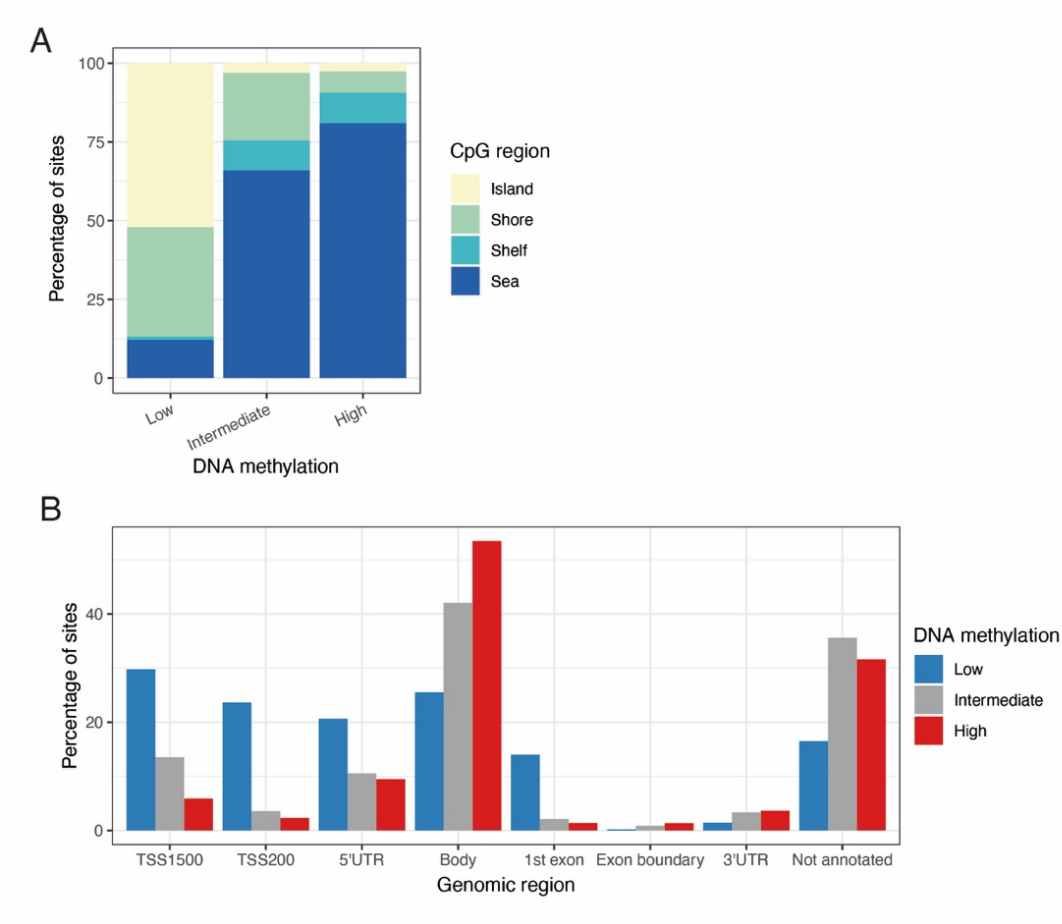


Figure 2. DNA methylation distribution of the human sperm DNA methylome.

A) The percentage of CpG sites that display low (median beta < 0.2), intermediate (median beta between 0.4 and 0.6) and high (median beta > 0.8) levels of DNA methylation in spermatozoa are shown according to CpG region.

B) The percentage of CpG sites that display low, intermediate and high levels of DNA methylation in spermatozoa are shown according to their genomic region.

TSS: transcription start site, UTR: untranslated region

DNA methylation in imprinted regions

Genomic imprinting refers to the phenomenon that genes are epigenetically regulated to be expressed in a parent-of-origin specific manner [23]. In spermatozoa, imprinted genes should be either completely unmethylated or fully methylated depending on the gene [23]. Conversely, in blood, the parent-of-origin driven allele-specific methylation should result in methylation values of around 50% for any given imprinted site. DNA methylation levels at CpG sites annotated to genes listed in the Geneimprint database (http://www.geneimprint.com/site/genes-by-species) were compared between spermatozoa and whole blood (**Figure S2**). In the case of CpG sites annotated to genes that are known to be imprinted, we observed an enrichment of sites with median methylation 0.5 in whole blood, particularly for paternally imprinted genes (21% sites with median beta between 0.4 and 0.6 vs 3% of sites across the array-wide background; $P < 1.00 \times 10^{-50}$, Fisher's exact test), but also for maternally imprinted genes (11% of sites; $P = 9.19 \times 10^{-9}$). For genes predicted to be imprinted according to the Geneimprint

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

database, there was a less pronounced enrichment (paternal: 6% of sites; $P = 0.01$; maternal: 6% of sites; $P = 0.04$). No such enrichment was observed for spermatozoal DNA methylation in any of the four categories ($P > 0.05$). Because gene annotation on the methylation array is based only on proximity, this approach includes many CpG sites not actually located in imprinting control regions (ICRs). Therefore, we also compared DNA methylation distributions at sites which specifically fall into known human ICRs as reported by WAMIDEX (https://atlas.genetics.kcl.ac.uk). This second approach further confirmed an enrichment of probes with around 50% methylation located in ICRs in blood compared to sperm (**Figure S3**). Strikingly, of the 169 CpG sites that fell into ICRs, the majority show median beta values around 0.5 (57% of sites with beta between 0.4 and 0.6, $P < 1.00 \times 10^{-50}$, Fisher's exact test vs array-wide background). On the other hand, nearly all of the 169 sites were completely unmethylated in sperm (94% with median beta < 0.2, $P < 1.00 \times 10^{-50}$).

The sperm DNA methylome exhibits a more polarised genome-wide DNA methylation profile than blood

We compared the overall distribution of DNA methylation levels across the blood and sperm genomes. Sperm displayed a more polarised methylation profile compared to blood, i.e. that both low and high median levels of methylation were more commonly seen in sperm (**Figure 3A**), with 33% of sites showing median beta < 0.2 in sperm vs 27% in blood and 49% of sites with median beta > 0.8 in sperm vs 35% in blood. Principal component (PC) analysis was performed across the full discovery dataset comprising the 704,356 probes that remained after filtering. The first PC, explaining 51.41% of the variance, clearly distinguished between sperm and blood, indicating that the tissue of origin was the primary determinant of differences in DNA methylation profiles (**Figure S4**). At the majority of interrogated sites, DNA methylation levels differed significantly between sperm and blood (n = 447,846 sites (64%), $P < 9 \times 10^{-8}$, paired t-test; **Table S3**). At 62% of these sites (n = 277,831 sites), sperm was relatively hypermethylated compared to blood.

A more detailed characterisation of the differences between the sperm and blood DNA methylomes was performed by comparing DNA methylation levels in sperm and blood across different genomic regions (**Figure 3B-C, Tables S5-S6**). CpG islands and CpG island shores were found to be less methylated in sperm compared to blood (0.07 and 0.16 lower in sperm respectively, $P < 1.0 \times 10^{-50}$ for both, paired t-test). CpG island shelves and CpG sites in open seas were relatively hypermethylated in sperm compared to blood (0.06 and 0.07 higher in sperm respectively, $P < 1.0 \times 10^{-50}$ for both) (**Figure 3B, Table S5**). Regions upstream of transcriptional start sites were relatively hypomethylated in sperm compared to blood (0.02 lower at TSS200 and 0.11 at TSS1500, $P < 1.0 \times 10^{-50}$ for both), as were sites mapping to the 3'UTR (0.01 lower, $P = 3.81 \times 10^{-5}$) or first exon (0.01 lower, $P < 1.0 \times 10^{-50}$). Conversely, other transcribed regions were hypermethylated in sperm compared to blood, including gene bodies (0.02 higher, $P < 1.0 \times 10^{-50}$), 5'UTRs (0.01 higher, $P = 1.3.61 \times 10^{-32}$), and exon boundaries (0.02 higher, $P = 2.80 \times 10^{-22}$; **Figure 3C, Table S6**). We replicated these differences in the lean replication (n = 21 lean males) and obesity cohort (n = 22 obese males) (**Supplementary Material: Replication, Figure S5, Table S3**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

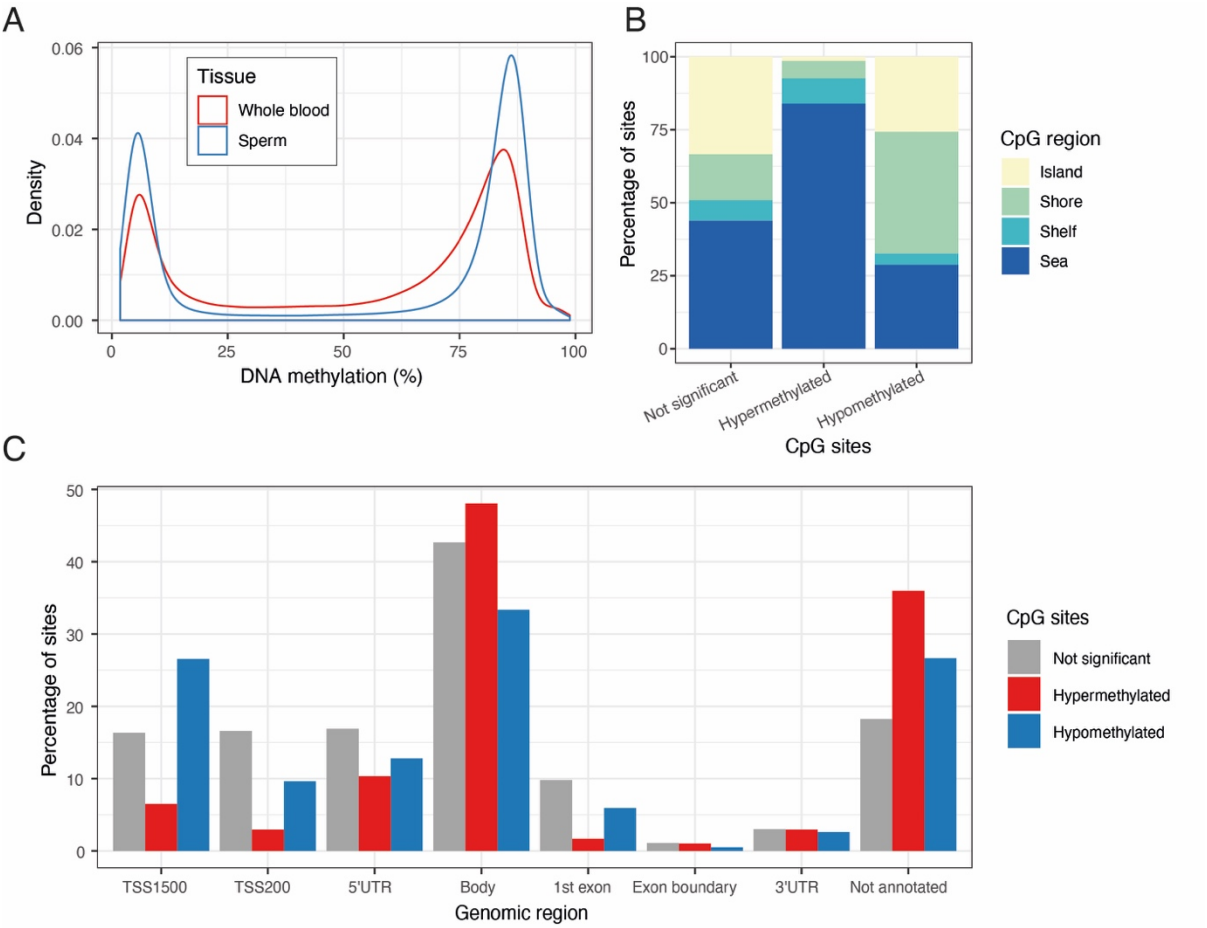


Figure 3. Comparison of DNA methylation levels in human sperm and whole blood.

A) Array-wide comparison of CpG methylation in sperm and blood, showing that both low (< 20%) and high (> 80%) DNA methylation levels are more commonly seen in sperm. Plotted is the distribution median DNA methylation levels across all individuals in the discovery cohort.

B) The percentage of CpG sites that are relatively hyper- and hypomethylated in sperm compared to blood, and CpG sites where there is no significant difference in DNA methylation between the tissues, are shown according to CpG region. **C)** The percentage of CpG sites that are relatively hyper- and hypomethylated in sperm compared to blood, and CpG sites where there is no significant difference in DNA methylation between the tissues, are shown according to genomic region.

TSS: transcription start site, UTR: untranslated region

Sperm has a unique DNA methylation profile enriched in pathways relating to transcriptional regulation

The Gene Expression Omnibus (GEO) is a publicly available data repository that contains DNA methylation data from a range of human tissue samples, most of which have been analysed using the Illumina Infinium HumanMethylation450 BeadChip (450K array) [21]. In order to investigate how the DNA methylation profile of spermatozoa compares to that of somatic tissues, DNA methylation data from 371 sperm samples (90 from our discovery, replication and obesity cohorts combined and 281 samples from GEO) was compared to that of 5,917 somatic tissue samples from male donors available on GEO (see **Table S7** and **Table S8** for details on tissue samples). Restricting analysis to CpG sites covered by both the EPIC and 450K arrays (n = 452,626 sites) we used linear regression to identify sperm-specific DNA methylation signals across the 6,288 samples. After Bonferroni correction, a total

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

of 133,125 genome-wide significant CpG sites (29%) were identified as differentially methylated between sperm and somatic tissues (**Table S9**). At 18% of these sites (n = 109,290 sites) sperm was characterized by higher methylation levels than somatic tissues. This is in contrast to the paired analysis with blood and likely due to the nearly exclusive coverage of CpG islands on the 450K array. Gene Ontology (GO) enrichment analysis [24] revealed 272 GO terms amongst hypermethylated CpG sites (**Table S10**). The main two categories of enriched pathways related to regulation of gene transcription (37 pathways) and neurological traits and functions (67 pathways). The latter is possibly driven by the relatively large proportion of brain and neuronal samples amongst the somatic tissues (16%). Of the 37 GO terms enriched amongst hypomethylated CpG sites, 8 (22%) related to sensory perception, particularly smell (**Table S11**). We repeated the same analysis removing unsorted tissues and tumours as well as cell lines (1,046 samples) and replicated virtually the same results.

Covariation of DNA methylation between sperm and blood is limited and most likely explained by genetic variation

We next explored whether, despite the blood and sperm DNA methylomes being highly distinct, there were CpG sites where the levels of DNA methylation covaried between the tissues. We used minimum variability criteria for sites to be tested to avoid correlations driven by individual outliers, similar to those used by Hannon and colleagues [15]: we selected sites for which the middle 80% of samples had a beta range ≥ 0.05 in both blood and sperm. This restricted our analyses to 155,269 variable sites. At 1,513 of these (~1%), DNA methylation levels were significantly correlated between the two tissues ($P < 9 \times 10^{-8}$, Pearson's product moment correlation; **Figure 4A, Table S12**).

Given the observation of several bi- and trimodal patterns of DNA methylation amongst highly correlated sites (**Figure 4B**), we applied a combination of outlier analysis and k means clustering with manual verification, to identify which of the 1,513 significantly correlated CpG sites exhibit these patterns. The majority of correlated CpG sites (1,140 sites, 75%) showed a bimodal distribution and 205 sites (14%) showed a trimodal distribution of DNA methylation, both of which are suggestive of a strong genetic influence on DNA methylation or the measurement. Probes with the highest correlation coefficients tended to show clear trimodal patterns (**Figure 4B**), while a third of bimodally distributed probes (365) appear to be driven by single outliers (**Figure S6**). A subset of correlated sites (30 i.e. 2%) displayed a negative correlation between DNA methylation in sperm and blood (**Figure 4C**) and at a small number of sites distinct trimodal methylation patterns are present in only one of the two tissues (**Figure 4D**).

We cross-checked all correlated sites for known SNPs in the probe sequence using the dbSNP Human Build 151 database [25]. Nearly all probes (1,507; > 99%) were found to have known SNPs in the probe sequence, > 90% of which are in the CpG site itself (**Figure 5**). This would indicate that DNA methylation readouts at these sites are most likely measuring genetic variation rather than epigenetic state. Only a small subset (n = 6) of the CpG sites that were significantly correlated had no known SNPs in their probe sequence. Some of these nevertheless displayed bi- and trimodal patterns of DNA methylation suggestive of a genetically driven effect and could potentially constitute strong mQTLs (**Figure 4E**).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

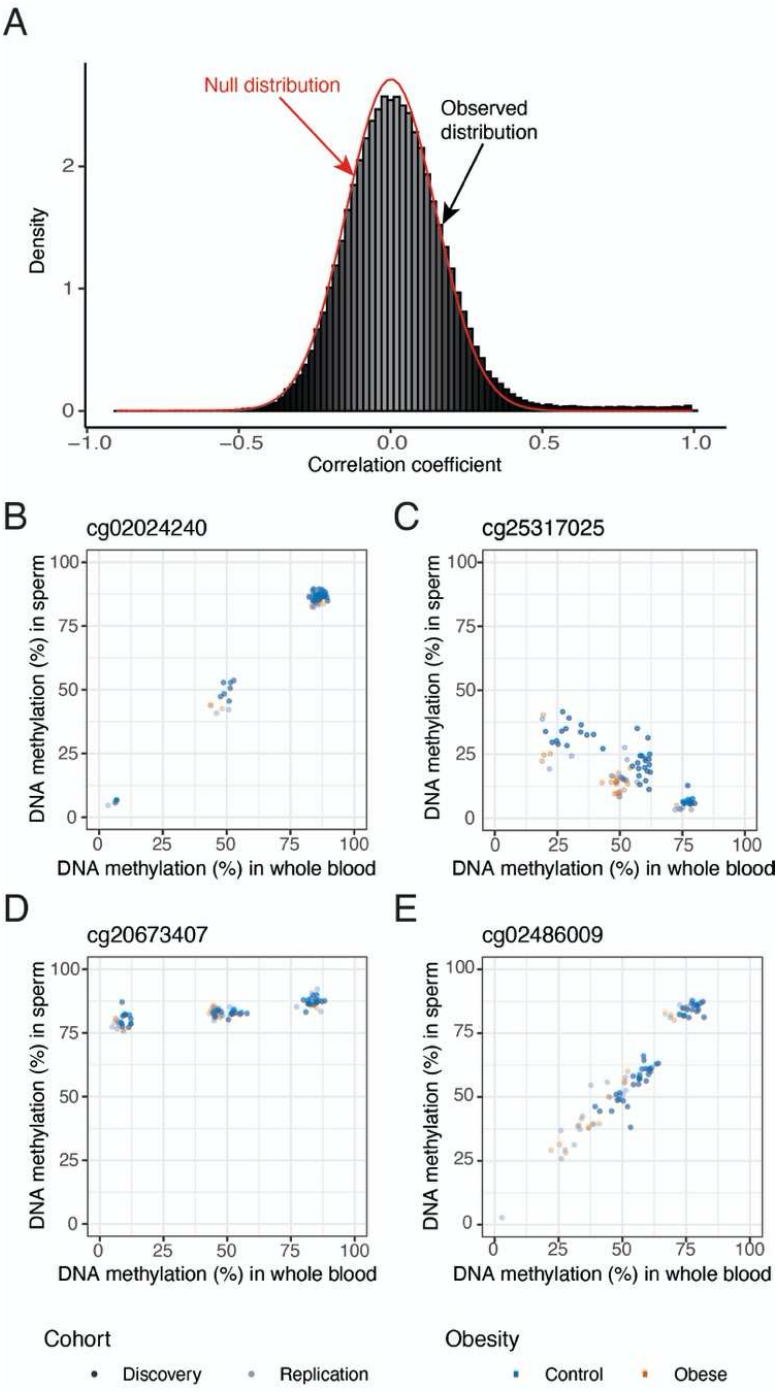


Figure 4. Covariation of DNA methylation between blood and sperm.

A) The observed correlation of DNA methylation levels in sperm and blood (histogram) is plotted against the estimated null distribution (red density curve). A small percentage of sites display highly correlated DNA methylation levels ($r > 0.8$), and the observed distribution is overall slightly shifted to the right compared to the null distribution.

B) cg02024240 (chr5:159669974) shows a strong DNA methylation correlation between blood and sperm and a trimodal methylation pattern suggestive of a genetically driven effect ($r > 0.99$, $P = 4.68 \times 10^{-48}$).

C) cg25317025 (chr18:47019823) is one of 30 sites showing a negative correlation between blood and sperm (r = -0.89, $P = 5.14 \times 10^{-17}$).

D) Some probes display striking differences in variability between the two tissues: cg20673407 (chr10:31040939) is characterized by a distinct trimodal pattern in whole blood while showing less overall variability in sperm ($r = 0.82$, $P = 1.45 \times 10^{-12}$).

E) Only 6 of the significantly correlated probes have no known SNPs anywhere in the probe sequence. cg02486009 (chr15: 22428395) is one of these ($r = 0.96$, P= 1.90×10^{-27}). Nonetheless it shows a bimodal DNA methylation pattern in both tissues, suggestive of a genetically driven effect.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

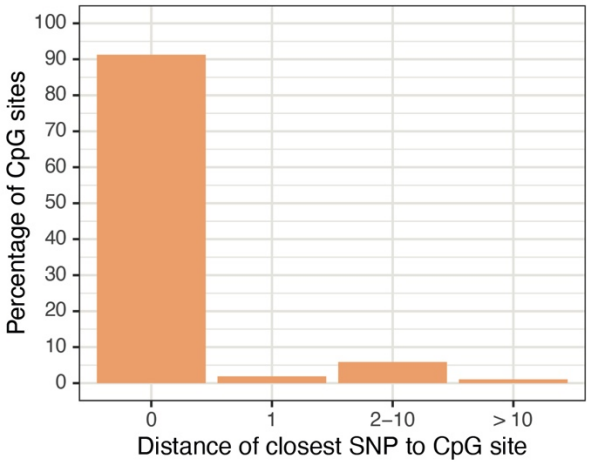


Figure 5. Positions of known SNPs in probe sequences of correlated probes.
1,507 of the 1,513 significantly correlated probes have known SNPs in their probe sequence. The vast majority of these (> 90%) map to the CpG site itself.

Secondly, we overlapped our correlated CpG sites with a list of recently reported correlated regions of systemic interindividual variation (CorSIV) in DNA methylation [26]. Only 0.2% of non-correlated variable probes are contained in CorSIVs – in line with the low overall genomic prevalence of these regions (0.1% of the human genome). Strikingly, we observe a 10-fold enrichment of this within the correlated sites (2.2%, $P = 8.85 \times 10^{-25}$, Fisher's exact test). The observations from the sperm data suggest that for sites exhibiting bi- and trimodal methylation patterns there is a likely genetic origin (of either a SNP in the CpG site or strong methylation QTL effects). Therefore, this enrichment conflicts with the hypothesis that for at least these sites, the origin of cross-tissue covariation is developmentally established stable epialleles [27]. Finally, using cis DNA methylation QTL data from whole blood published by McClay and colleagues [28] we found that 232 (30%) of the correlated sites also present on the 450K array had previously been identified as mQTLs in whole blood, representing a significant enrichment over the 16% observed across all variable probes ($P = 1.66 \times 10^{-33}$, Fisher's exact test). Correlations largely replicated in the two replication cohorts. (**Supplementary Materials: Replication**, **Table S12**) and non-replicating sites were generally driven by outliers in the discovery cohort (examples shown in **Figure S7**).

Limited evidence for converging associations between DNA methylation and obesity from whole blood and sperm

We next investigated whether obesity was associated with DNA methylation in sperm or blood. At the 697,384 sites that passed quality control in the combined replication cohort, including lean and obese males, we used linear regression of DNA methylation on obesity status, controlling for estimated blood cell types in the blood dataset. No probes passed array-wide significance ($P < 9 \times 10^{-8}$) in blood or sperm (**Table S13**). Given our small sample size, we leveraged published data from a larger EWAS of BMI in whole blood [1]; see **Materials and Methods**). First, we tested whether the 187 replicated array-

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

wide significant probes ($P < 1.0 \times 10^{-7}$) reported by Wahl and colleagues, which were also present in our data, were enriched in lower-ranked P values in our data, and secondly, we compared effect sizes at these 187 probes between our cohort and the published data. To make both analyses comparable we treated BMI as a continuous measure for these comparisons – as Wahl and colleagues had done in the original epigenome-wide association study. Both analyses confirmed enrichments of the reported associations in blood but not sperm: lower-ranked P values were enriched in blood ($P < 1.3 \times 10^{-23}$, Wilcoxon rank sum test) but not sperm ($P = 0.06$, **Figure 6A**) and similarly, the reported effects at the 187 probes were correlated significantly with effects observed in our blood data ($\rho = 0.72$, $P < 1.0 \times 10^{-50}$, Spearman's rank correlation, **Figure 6B**) but not in sperm ($\rho = 0.13$, $P = 0.11$, **Figure 6C**). This indicates that the associations identified by Wahl and colleagues do not generalize to sperm. Finally, to maximise power within our own sample, we ran a linear mixed effects model across the discovery and replication datasets, using the 692,265 probes that survived quality control in both datasets. DNA methylation was regressed onto tissue (blood versus sperm), age, batch and obesity status, while controlling for interindividual variation with a random effect (**Table S13**). This analysis found that methylation at one CpG site, cg19357369 (chr4:2429884), was significantly increased in obese men in sperm and blood (beta = 0.02, $P = 8.95 \times 10^{-8}$, **Figure 6D**).

Obesity does not significantly influence the covariation of DNA methylation between sperm and blood

To investigate whether the covariation of DNA methylation was significantly altered in obesity, we ran an interaction model that regressed DNA methylation in blood onto DNA methylation in sperm, obesity status and their interaction effect, while covarying for experimental batch and age (see **Materials and Methods**). We identified 98 CpG sites with a statistically significant interaction between obesity and the association of blood and sperm DNA methylation ($P < 9 \times 10^{-8}$). Interactions at the vast majority of these CpG sites (96) were driven by individual outliers in the obese cohort (**Figure S8A-C**); the remaining two sites appear to be driven by outliers in the lean cohort and a batch effect (**Figure S8D**). We therefore conclude that we were not able to identify credible altered DNA methylation covariation patterns between blood and sperm that may have arisen as part of a gene-environment interaction.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

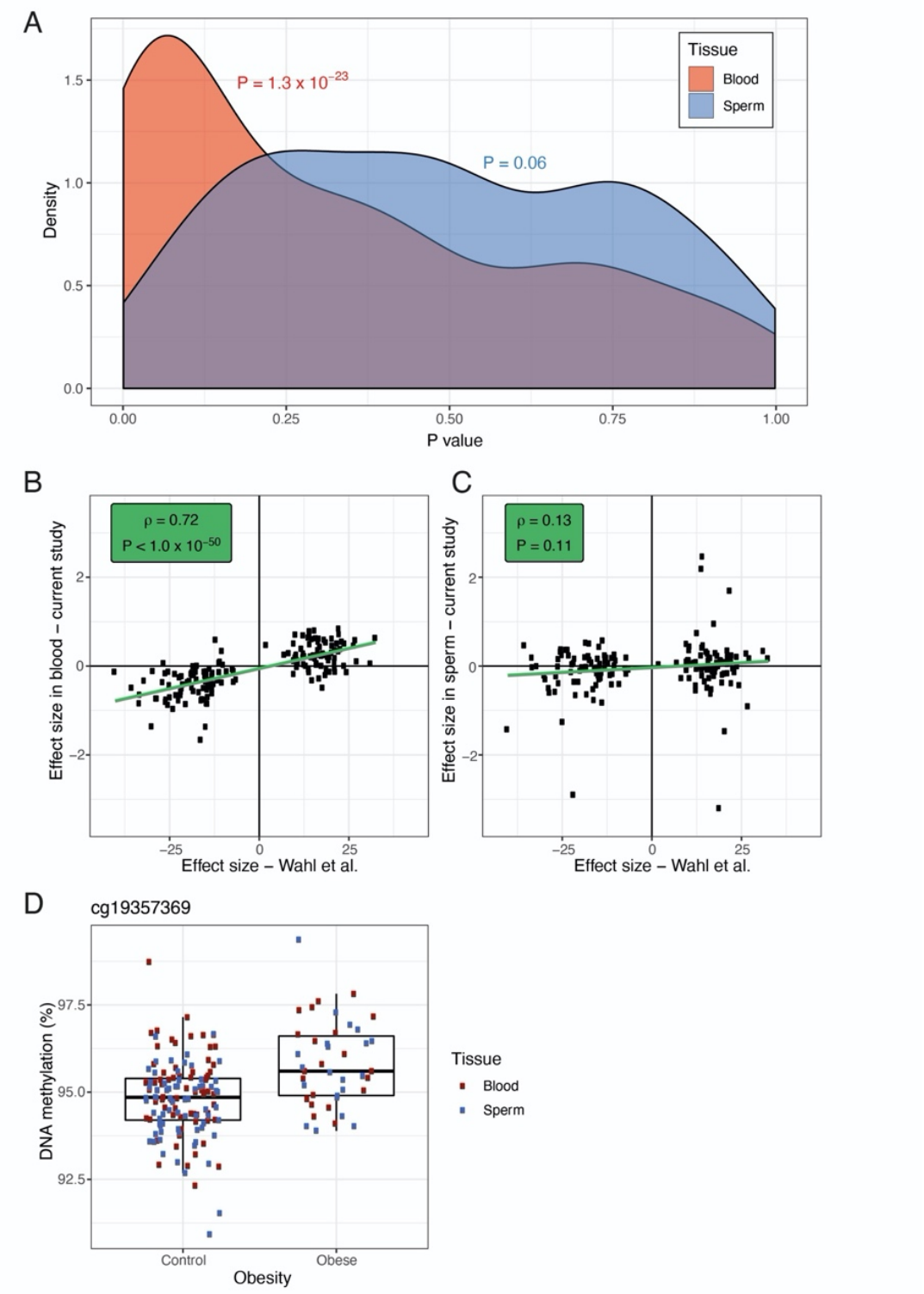


Figure 6. Obesity associated DNA methylation patterns in whole blood and sperm.

Out of all replicated CpG sites reported to be associated with BMI by Wahl et al. ($P < 1.0 \times 10^{-7}$), 187 were also present in our replication cohort of lean and obese men. We regressed BMI onto DNA methylation in each tissue, controlling for estimated blood cell types in the blood analysis to match the analysis used by Wahl and colleagues. **A)** Lower-ranked P values were found to be enriched amongst these 187 sites in blood ($P < 1.3 \times 10^{-23}$, Fisher's exact test) but not sperm ($P = 0.06$). **B)** Effect sizes at the 187 probes were significantly correlated between our blood data and the summary statistics published by Wahl and colleagues ($\rho = 0.72$, $P < 1.0 \times 10^{-50}$, Spearman's rank correlation). **C)** No such correlation was observed for our sperm data ($\rho = 0.13$, $P = 0.11$). **D)** In a linear mixed effects model across the discovery and replication datasets, DNA methylation was regressed onto tissue (blood versus sperm), age, batch and obesity status, while controlling for interindividual variation. This analysis identified significant hypermethylation at one CpG site, cg19357369 (chr4:2429884), in obese compared to lean men across the two tissues (beta difference = 0.02, $P = 8.95 \times 10^{-8}$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

In this study, we characterized the sperm methylome in relation to blood and other somatic tissues, investigated covariation between DNA methylation in sperm and whole blood and analyzed DNA methylation patterns associated with obesity. We conclude that the DNA methylation profiles of sperm and blood are highly distinct, and that there is little evidence of DNA methylation covariation between the two tissues, beyond genetic and technical effects.

In line with previous, smaller-scale studies, we showed that the sperm DNA methylome is highly polarised compared to that of blood, with both low (beta < 0.2) and high (beta > 0.8) levels of DNA methylation more frequently observed in sperm than in blood [20]. In contrast to previous research, however, we found that the sperm DNA methylome is overall slightly hypermethylated compared to that of blood [20, 29, 30]. This finding is potentially influenced by the fact that the previous generations of DNA methylation arrays (the 450K array) included a higher proportion of CpG islands, which are relatively hypomethylated in spermatozoa [20, 31].

We identified significant differences in DNA methylation levels at the majority of assayed CpG sites when comparing whole blood to sperm. Additionally, in our comparison of the spermatozoal DNA methylome to that of almost 6,000 somatic tissue samples, we showed that gene ontology terms enriched amongst hypermethylated CpG sites in sperm pointed repeatedly to transcriptional regulation. This is an intriguing finding considering that recent research has shown that high overall levels of transcription during spermatogenesis facilitate transcription-coupled DNA repair mechanisms through so-called “transcriptional scanning” [32]. Given that transcriptional regulation is an essential process for all cell-types, it is striking to observe sperm-specific DNA methylation patterns enriched in these processes. It could suggest that DNA methylation is involved in widespread transcriptional downregulation as cells progress from an active transcriptional stage during spermatogenesis to a more transcriptionally repressed stage in mature sperm.

About 1% of variable sites in whole blood and sperm showed a significant correlation of DNA methylation between the whole blood and sperm. This is slightly lower than what has been reported for comparisons of DNA methylation between whole brain and peripheral tissues [33]. Furthermore, at the vast majority of correlated CpG sites, the correlation appeared to be driven by underlying genetic variation resulting in characteristic bi- and trimodally clustered distributions of DNA methylation. In most of these cases, known SNPs were identified in the CpG site itself or in the single base extension. This finding is further supported by the observed enrichment of mQTLs [28] and CorSIVs [26] amongst correlated sites. Thus, whilst we lack specific genotyping information on individual participants in this study, our findings strongly suggest genetic variation as the underlying cause of DNA methylation covariation between blood and sperm. This is despite the fact that we employed stringent filtering of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

probes in close proximity to SNPs from previously published lists [31, 34, 35], which suggests a need to update existing reference lists.

We also identified a small number of CpG sites where DNA methylation was negatively correlated between blood and sperm, and sites where DNA methylation exhibited a trimodal distribution pattern in one tissue only. It would be of interest to investigate further whether pathophysiological traits are associated with an increase in DNA methylation in one tissue and a decrease in the other. In particular, whether germ cell or leukocyte specific transcription factors are responsible for the discordant yet correlated DNA methylation distribution patterns across blood and sperm.

The small number of sites (6 out of 1,513) where no obvious genetic driver of methylation variability was identified are likely too few to be of value in studies where blood is needed as a surrogate tissue for sperm. The results of this study are generally in line with similar studies of DNA methylation covariation, such as between whole blood and various brain regions [15], albeit more extreme. They emphasize the importance of using disease-relevant tissues in epigenomic investigations. These findings do not however, generally preclude the use of readily accessible tissues such as blood or saliva for identifying DNA methylation biomarkers of conditions relating to germ cell function, such as subfertility. For example, if a robust DNA methylation profile of subfertility is identified in blood, this could be a helpful test in fertility evaluations without necessarily reflecting the epigenetic profile of spermatozoa.

This study identified one CpG site, cg19357369, as hypermethylated in sperm and blood from obese versus lean males. The finding should be interpreted with caution as it requires replication and just passed the array-wide multiple testing threshold – which was not corrected for the different aspects pertaining to sperm DNA methylation across the study (comparison with blood, correlation with blood, interaction, single-tissue EWAS, multi-tissue EWAS). The effect size was also comparatively small (beta = 0.02). cg19357369 is found upstream of the lncRNA *RP11-503N18,* which has yet to be characterised in terms of biological function [36]. However, previous research has shown that DNA methylation at cg19357369 is significantly altered during human fetal brain development [37]. Although cg19357369 has previously been identified as differentially methylated in hepatic tissue from obese compared to lean males lean males [36], it has not previously been identified in EWASs of obesity or BMI when only blood samples have been analysed. If shown to be replicable, it could point towards the possibility of an obesity associated signature of spermatozoa.

Overall, we found that differentially methylated CpG sites associated with BMI in a large-scale EWAS in blood were not evident in sperm. Therefore, our current understanding of epigenetic associations of weight-associated phenotypes, which stems almost exclusively from studies of whole blood, is unlikely to give us functional insights into how these may be passed to offspring.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

There are limitations to our study. First, it constitutes an observational, cross-sectional study and we are therefore unable to comment on the causality behind observed associations between obesity and spermatozoal DNA methylation. The limited sample size of the obesity cohort (n = 22) reduced our ability to detect modest effects of obesity on DNA methylation covariation between sperm and whole blood. The obesity cohort included a proportion of overweight males (BMI 25-30 kg/m^2), which potentially diluted our results. Further, while we used the most comprehensive DNA methylation array currently available, the MethylationEPIC array is still biased towards certain parts of the genome (most notably enhancer regions, RefSeq genes and CpG islands) and does not give a complete picture of genome-wide CpG methylation [38]. Lastly, although we were able to speculate as to the effects of genetic variants in CpG sites influencing our results, given trimodal methylation patterns and the presence of known SNPs in the CpG site, we did not have the actual genetic sequence of our subjects to verify this directly.

The study has several strengths. It constitutes the largest unbiased analysis of DNA methylation in matched human sperm and blood samples performed to date, and is one of the largest studies of spermatozoal DNA methylation in healthy males of proven fertility. In contrast to several previous analyses of DNA methylation in human spermatozoa [39-41], our study includes a replication cohort, increasing the robustness of our findings. Crucially, our analyses include the use of large existing datasets; blood-sperm correlated CpG sites were interrogated for overlap with previously identified mQTLs in whole blood [28] as well as with a list of recently reported CorSIVs [26]. We used findings from one of the largest studies of obesity-associated DNA methylation in blood performed to date [1] to analyse whether effects of obesity observed in blood overlapped with those observed in sperm. Lastly, we used recently developed DNA methylation analysis pipelines for large DNA methylation datasets [42] to identify sperm-specific DNA methylation signatures by comparing spermatozoal DNA methylation data to that of almost 6,000 somatic tissue samples available on GEO [21]. Together, these analyses allowed us to interrogate the spermatozoal DNA methylome in novel ways and provide highly suggestive evidence for why DNA methylation as a mechanism for intergenerational effects of obesity in humans is unlikely.

Conclusions

Our data suggests that compared with a wide range of somatic tissues, human sperm displays a unique DNA methylation profile, particularly in pathways relating to transcriptional regulation. We show that DNA methylation levels in human blood and sperm are only correlated at a minority of CpG sites and that at such sites, DNA methylation covariation is most likely due to genetic effects. The use of peripheral blood as a surrogate tissue for human spermatozoa is therefore inadvisable. Obesity does not generally influence spermatozoal DNA methylation, nor the covariation of DNA methylation between blood and sperm. Further, obesity-associated CpG sites identified in peripheral blood do not show enrichment in spermatozoa from obese individuals. Taken together, our findings suggest that if there

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

are inter- and transgenerational effects of human obesity, they are unlikely to be mediated by changes in spermatozoal DNA methylation.

Materials and Methods

Samples

Whole blood and semen samples were collected from participants recruited from University College London Hospital (UCLH) May 2016 - March 2019. Participants were phenotyped with regards to BMI, waist circumference, systolic and diastolic blood pressure, blood lipids, fasting insulin and glucose levels and C-reactive protein (CRP). Phenotypic information about participants is detailed in **Table S4**. Participants provided information about their medical history and lifestyle via questionnaires, and were excluded if they suffered from significant medical conditions or took regular medications. All participants were of proven fertility. Peripheral blood samples were centrifuged at 3000g for 15 minutes within one hour of venepuncture and the buffy coat was used for DNA extraction.

Semen samples were processed within one hour of sample production as per UCLH protocol and analysed for sperm concentration, motility and average progressive velocity using the Sperminator/Computer Assisted Sperm Analysis system (Pro-Creative Diagnostics, Staffordshire, UK). Semen sample parameters are detailed in **Table S14**. All semen samples were within normal parameters according to World Health Organization criteria [43]. Samples underwent gradient centrifugation (45 and 90% PureSperm medium; PureSperm 100®, Nidacon Laboratories, PS100-100) to select for motile spermatozoa as described elsewhere [44]. The processed samples were microscopically assessed for cell purity such that only samples with no visible cells other than spermatozoa were included in downstream analyses.

DNA extraction

DNA from 200 μL buffy coat derived from whole blood was extracted using Qiagen QIAamp DNA Blood Mini Kit (Qiagen, Cat No. 51104) according to manufacturer's instructions [45]. DNA from the pellet of motile spermatozoa was extracted using a standard phenol-chloroform extraction method as described previously [46]. DNA extracted from whole blood and sperm was quality controlled using a Qubit 3.0 Fluorometer (Life Technologies, Cat No. Q33216). DNA was stored in -80°C prior to bisulphite conversion.

Methylomic profiling

DNA (500 ng) from each sample was sodium bisulphite-treated using the Zymo EZ 96 DNA methylation kit (Zymo Research, Cat No. D5004) according to the manufacturer's instructions. DNA methylation was quantified using the Illumina Infinium MethylationEPIC BeadChip [38] using an Illumina iScan System [47]. Samples were assigned a unique code for identification and randomized with regards to cohort and other variables to avoid batch effects, and processed in two batches. The Illumina Genome

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Studio software was used to extract the raw signal intensities of each probe (without background correction or normalization). Raw DNA methylation data is available for download from GEO (accession number GSE102538).

Data pre-processing

Data analysis was performed in R version 3.6.2. DNA methylation data was processed and analysed using the *wateRmelon* package in R [48]. An initial outlier analysis was performed using the outlyx() function in *wateRmelon* based on 1) the interquartile range of the first principal component and 2) the pcout algorithm [50] detecting outliers in high dimensional datasets, leading to the removal of 1 individual from the discovery cohort, 2 individuals from the obesity cohort and 3 Individuals from the lean replication cohort. The 59 non-CpG SNP probes on the array were used to confirm that the genotypes at these 59 probes were identical for the matched samples.

Prior to data analysis, 9,779 probes were removed from the discovery data because more than 5% samples displayed a detection P value > 0.05. Furthermore, 3,337 probes were removed because of having a bead count < 3. Probes containing SNPs in close proximity to the CpG site (within 10 base pairs) as well as potentially cross-reactive probes were filtered using annotated lists from three sources [31, 34, 35], leading to the removal of 149,105 CpG sites. The final discovery data set comprised 704,356 CpG sites. Data was normalized in the R package *wateRmelon* using the dasen() function as previously described [48]. The lean and obese replication cohort were processed together experimentally and therefore jointly pre-processed and normalised using the same parameters as for the discovery dataset. A total of 697,442 probes survived quality control and filtering in the replication data. DNA methylation was analysed and reported as beta values, which is the ratio of methylated probe intensity over the overall intensity and approximately equal to the percentage of methylated sites (% DNA methylation). For plotting purposes, beta values are shown and described and shown as percent DNA methylation.

Data analysis

Characterization of DNA methylation in sperm

CpG sites were assigned to chromosomes, locations, genes, and genomic regions using the Illumina manifest for the EPIC array (hg19 reference). CpG sites were classified as having either 'high' (median beta > 0.8) or 'low' (median beta < 0.2) DNA methylation. Enrichments of each genomic or CpG region amongst 'high' and 'low' methylation sites were calculated against the background (sites showing 0.2-0.8 median beta values) using a Fisher's exact test.

DNA methylation age estimates

DNA methylation age was estimated on the discovery sample from both blood and sperm DNA methylation using Horvath's DNA methylation age estimator [4]. We additionally estimated DNA methylation age from sperm using the method described by Jenkins and colleagues [22].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Annotation of imprinted genes/ imprinting control regions

CpG sites were annotated to imprinted genes using the Illumina manifest for the EPIC array and the list of imprinted genes published in the Geneimprint database (http://www.geneimprint.com/site/genes-by-species). Enrichments of intermediate methylation levels were calculated as Fisher's exact tests of number of sites with median beta levels between 0.4 and 0.6 annotated to imprinted genes against the array-wide background. For known human imprinting control regions (ICR) we used the locations reported by WAMIDEX (https://atlas.genetics.kcl.ac.uk), these were lifted to hg19 and overlapped with CpG locations using the R package *GenomicRanges* [51]. Enrichments for intermediately methylated (median beta between 0.4 and 0.6) and unmethylated (median beta < 0.2) sites were calculated as Fisher's exact tests.

DNA methylation differences between blood and sperm

Sites characterized by differences in DNA methylation between whole blood and sperm were identified by a paired t-test of matched samples. Comparison of the difference in DNA methylation levels between sperm and blood at different genomic regions was performed by calculating a paired t-test of median DNA methylation in sperm vs blood across all sites annotated to a specific genomic or CpG region.

GEO analysis

DNA methylation data for 6,288 samples was downloaded from the Gene Expression Omnibus (GEO) including 281 sperm samples and 5,971 somatic tissue samples from male donors, profiled using the 450K or EPIC arrays. Statistical analyses were performed using the *bigmelon* package in R and statistical tests were performed using *limma* [42, 49]. In the comparison of DNA methylation between sperm and tissue samples from males on GEO, a linear model was fitted using the lmFit() function from the *limma* R package [49] across the 452,626 CpG sites that are present on both the EPIC and 450K arrays. The model regressed DNA methylation onto tissue (sperm vs not sperm) and included age and array type (450K or EPIC) as covariates. For sperm samples from GEO which lacked recorded age, the estimated age based on Jenkin's model was used instead. The data was not normalised because global large-scale differences between somatic tissues and sperm were expected, and because the high number of different types of samples included was expected to ameliorate issues around technical noise. The gene ontology (GO) pathway analysis was performed using the gometh() function from the *missMethyl* R package [52], which removes ambiguously assigned probes from the enrichment analysis.

Correlation between whole blood and sperm DNA methylation

In order to minimise the effect single outliers would have on the correlation analysis, a subset of 'variable' probes was identified by calculating the DNA methylation difference between the 10th and 90th percentile across all samples, and selecting sites where this was at least 0.05 in both whole blood and sperm (n = 155,269 sites). This approach is similar to the one described by Hannon and colleagues

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

previously [15]. Correlated CpG sites between sperm and blood were identified by Pearson's correlation test across all variable probes. In order to establish the matching null distribution, samples were permuted 100 times and correlations between DNA methylation in whole blood and sperm were recalculated across all variable sites. The density curve of these simulated correlations was added to the histograms of the empirical correlation coefficients to represent the null distribution (**Figure 4**). To investigate the clustering of DNA methylation patterns at significantly correlated CpG sites, a two dimensional outlier test was used by adapting the rosnerTest() function from the *EnvStats* R package [53] to exclude unimodal distributions. Next, k means clustering was applied for 2 and 3 clusters as implemented in the function pamk() of the R package *cluster* [54]. This function determines the best fitting number of clusters (two or three – corresponding to bi- and tri-modal methylation distributions). We manually checked and, if necessary, reassigned clusters which exhibited low between-cluster to within-cluster variance ratios (ratio < 2).

Annotation of SNPs and genetic enrichments

To annotate SNPs to their location within probe sequences we used the Illumina EPIC hg38 manifest and dbSNP database build 151 in the *SNPlocs.Hsapiens.dbSNP151.GRCh38* R package. SNPs were mapped to probes using the *GenomicRanges* R package [51] and the distance to the CpG site of the closest SNP in the probe sequence was calculated for each of the 1,513 probes with significant correlations between sperm and blood. We downloaded the locations of the 9,226 correlated regions of systemic interindividual variation (CORSIV) in DNA methylation recently published by Gunasekara and colleagues [26]. These were overlapped with the locations of CpG sites using the hg38 manifest and the *GenomicRanges* R packages. Finally, we downloaded the list of cis methylation QTLs (mQTLs) in blood reported by McClay and colleagues [28]. These were identified using the 450K array, which meant we had to restrict this annotation to probes present on both the EPIC and 450K array. Enrichments for CORSIVs and mQTLs were calculated by Fisher's exact test against the background of non-correlated variable probes.

Obesity and DNA methylation in blood and sperm

Two models were used to investigate the association between obesity and DNA methylation in sperm and blood. First, DNA methylation was regressed onto obesity status in the combined replication cohort, in blood and sperm separately. This analysis was controlled for estimated blood cell counts in blood. Secondly, a mixed effects model was run across both the discovery and replication cohorts using the lmer() function from the *lme4* package in R [55], regressing DNA methylation onto tissue (blood versus sperm), age, batch and obesity status, while controlling for interindividual variation with a random effect:

lmer(Methylation ~ Tissue + Age + Batch + Obesity +(1|ID))

Given our small sample size – especially in the obese group - we downloaded summary statistics from an EWAS of BMI in whole blood [1]. 187 of the replicated array-wide significant probes ($P < 1.0 \times 10^{-7}$) reported by Wahl and colleagues were also present in our dataset. To make our data comparable we treated BMI as a continuous measure for these comparisons, regressing BMI onto obesity status and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

controlling for estimated blood cell proportions in the blood analysis. We tested for an enrichment of lower ranked P values amongst the 187 previously reported probes in our analysis using a Wilcoxon rank sum test. Secondly, we looked at correlations of effect sizes reported by Wahl and colleagues and observed in our data across the 187 probes using Spearman's rank correlation to allow for study-specific biases.

Interaction between obesity, tissue and DNA methylation

To detect and interaction between obesity and the association between blood and sperm DNA methylation we ran linear model regressing DNA methylation in blood onto DNA methylation in sperm, obesity status and their interaction effect, while covarying for experimental batch and age:

$lm(Methylation_{Blood} \sim Methylation_{sperm} * Obesity + Age + Batch)$

Cell-type composition

As whole blood represents a heterogenous tissue where the composition of leukocytes can introduce bias in the interpretation of DNA methylation analysis findings, blood cell type counts of monocytes, granulocytes, NK-cells, B cells, CD8+-T-cells, and CD4+-T-cells were estimated from the DNA methylation data using the method described by Houseman [56]. These estimates were included in all analyses that were run on the blood dataset alone as described above.

Multiple testing correction

For agnostic analyses across the whole EPIC array (including those restricted to variable probes), the threshold $P < 9 \times 10^{-8}$ as reported in recently published statistical guidelines for the EPIC array [57]. For the GEO analysis only the set of probes present on both the 450K and EPIC array were used. We applied Bonferroni correction across these 452,626 sites.

Declarations

Ethics approval and consent to participate

Ethical approval for the study was granted from the South East Coast - Surrey Research Ethics Committee on 28 September 2015 (REC reference number 15/LO/1437, IRAS project ID 164459). The study was also registered with the University College London Hospital Joint Research Office (Project ID 15/0548). All participants provided written, informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the Gene Expression Omnibus repository, under GEO accession number GSE149318.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Funding

FA was supported by a studentship from the Rosetrees Trust (Ref No A815) and the work was supported by a Medical Research Council grant (MRC reference code MR/P011799/1). SJM is funded by an Edmond and Lily Safra Research Fellowship and a UK Dementia Research Institute Career Development Fellowship. The UK Dementia Research Institute receives its funding from UK DRI Ltd, funded by the UK Medical Research Council, Alzheimer's Society and Alzheimer's Research UK. DJW is supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

Authors' contributions

SJM, FA, DJW and VKR designed the study. DJW and VKR acquired funding and SJM supervised the study. FA recruited participants, collected samples and extracted DNA with advice from MLH. EW oversaw the processing on semen samples. FA and AB bisulphite converted DNA. YP processed the EPIC Array. SJM carried out data pre-processing and bioinformatics analyses. TJGS performed the comparison of sperm DNA methylation to tissues on GEO. SJM, FA, LCS, DJW and VKR contributed towards the interpretation of the data. FA and SJM drafted the manuscript. MLH, DJW, VKR, LCS and TJGS all made critical revisions. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank the technicians at UCL Genomics at the Great Ormond Street Institute of Child Health for processing of the Infinium MethylationEPIC Array, Anna Greco for her role in recruiting participants, and Dr Sara Hillman and Dr Rob Lowe for previous work and helpful discussions on DNA methylation studies of obesity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC, Ried JS, Zhang WH, Yang YW, et al: **Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity.** *Nature* 2017, **541:**81-+.
2. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, Guan WH, Xu T, Elks CE, Aslibekyan S, et al: **Epigenetic Signatures of Cigarette Smoking.** *Circulation-Cardiovascular Genetics* 2016, **9:**436-447.
3. Mendelson MM, Marioni RE, Joehanes R, Liu CY, Hedman AK, Aslibekyan S, Demerath EW, Guan WH, Zhi DH, Yao C, et al: **Association of Body Mass Index with DNA Methylation and Gene Expression in Blood Cells and Relations to Cardiometabolic Disease: A Mendelian Randomization Approach.** *Plos Medicine* 2017, **14**.
4. Horvath S: **DNA methylation age of human tissues and cell types.** *Genome Biology* 2013, **14**.
5. Barbosa TD, Ingerslev LR, Alm PS, Versteyhe S, Massart J, Rasmussen M, Donkin I, Sjogren R, Mudry JM, Vetterli L, et al: **High-fat diet reprograms the epigenome of rat spermatozoa and transgenerationally affects metabolism of the offspring.** *Molecular Metabolism* 2016, **5:**184-197.
6. Sakai K, Ideta-Otsuka M, Saito H, Hiradate Y, Hara K, Igarashi K, Tanemura K: **Effects of doxorubicin on sperm DNA methylation in mouse models of testicular toxicity.** *Biochemical and Biophysical Research Communications* 2018, **498:**674-679.
7. Dias BG, Ressier KJ: **Parental olfactory experience influences behavior and neural structure in subsequent generations.** *Nature Neuroscience* 2014, **17:**89-96.
8. Watkins AJ, Dias I, Tsuro H, Allen D, Emes RD, Moreton J, Wilson R, Ingram RJM, Sinclair KD: **Paternal diet programs offspring health through sperm- and seminal plasma-specific pathways in mice.** *Proceedings of the National Academy of Sciences of the United States of America* 2018, **115:**10064-10069.
9. Youngson NA, Lecomte V, Maloney CA, Leung P, Liu J, Hesson LB, Luciani F, Krause L, Morris MJ: **Obesity-induced sperm DNA methylation changes at satellite repeats are reprogrammed in rat offspring.** *Asian Journal of Andrology* 2016, **18:**930-936.
10. Radford EJ, Ito M, Shi H, Corish JA, Yamazawa K, Isganaitis E, Seisenberger S, Hore TA, Reik W, Erkek S, et al: **In utero undernourishment perturbs the adult sperm methylome and intergenerational metabolism.** *Science* 2014, **345:**785-+.
11. Huypens P, Sass S, Wu M, Dyckhoff D, Tschop M, Theis F, Marschall S, de Angelis MH, Beckers J: **Epigenetic germline inheritance of diet-induced obesity and insulin resistance.** *Nature Genetics* 2016, **48:**497-+.
12. Wei YC, Yang CR, Wei YP, Zhao ZA, Hou Y, Schatten H, Sun QY: **Paternally induced transgenerational inheritance of susceptibility to diabetes in mammals.** *Proceedings of the National Academy of Sciences of the United States of America* 2014, **111:**1873-1878.
13. Tang WWC, Dietmann S, Irie N, Leitch HG, Floros VI, Bradshaw CR, Hackett JA, Chinnery PF, Surani MA: **A Unique Gene Regulatory Network Resets the Human Germline Epigenome for Development.** *Cell* 2015, **161:**1453-1467.
14. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al: **Integrative analysis of 111 reference human epigenomes.** *Nature* 2015, **518:**317-330.
15. Hannon E, Lunnon K, Schalkwyk L, Mill J: **Interindividual methylomic variation across blood, cortex, and cerebellum: implications for epigenetic studies of neurological and neuropsychiatric phenotypes.** *Epigenetics* 2015, **10:**1024-1032.
16. Soubry A, Murphy SK, Wang F, Huang Z, Vidal AC, Fuemmeler BF, Kurtzberg J, Murtha A, Jirtle RL, Schildkraut JM, Hoyo C: **Newborns of obese parents have altered DNA methylation patterns at imprinted genes.** *International Journal of Obesity* 2015, **39:**650-657.
17. Oldereid NB, Wennerholm UB, Pinborg A, Loft A, Laivuori H, Petzold M, Romundstad LB, Soderstrom-Anttila V, Bergh C: **The effect of paternal factors on perinatal and paediatric outcomes: a systematic review and meta-analysis.** *Human Reproduction Update* 2018, **24:**320-389.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18. McCowan LME, North RA, Kho EM, Black MA, Chan EHY, Dekker GA, Poston L, Taylor RS, Roberts CT: **Paternal Contribution to Small for Gestational Age Babies: A Multicenter Prospective Study.** *Obesity* 2011, **19:**1035-1039.
19. Tyrrell JS, Yaghootkar H, Freathy RM, Hattersley AT, Frayling TM: **Parental diabetes and birthweight in 236 030 individuals in the UK Biobank Study.** *International Journal of Epidemiology* 2013, **42:**1714-1723.
20. Krausz C, Sandoval J, Sayols S, Chianese C, Giachini C, Heyn H, Esteller M: **Novel Insights into DNA Methylation Features in Spermatozoa: Stability and Peculiarities.** *Plos One* 2012, **7**.
21. Clough E, Barrett T: **The Gene Expression Omnibus Database.** *Statistical Genomics: Methods and Protocols* 2016, **1418:**93-110.
22. Jenkins TG, Aston KI, Cairns B, Smith A, Carrell DT: **Paternal germ line aging: DNA methylation age prediction from human sperm.** *Bmc Genomics* 2018, **19**.
23. Barlow DP, Bartolomei MS: **Genomic Imprinting in Mammals.** *Cold Spring Harbor Perspectives in Biology* 2014, **6**.
24. Carbon S, Dietze H, Lewis SE, Mungall CJ, Munoz-Torres MC, Basu S, Chisholm RL, Dodson RJ, Fey P, Thomas PD, et al: **Expansion of the Gene Ontology knowledgebase and resources.** *Nucleic Acids Research* 2017, **45:**D331-D338.
25. **dbSNP Human Build 151 database** [https://www.ncbi.nlm.nih.gov/snp/]
26. Gunasekara CJ, Scott CA, Laritsky E, Baker MS, MacKay H, Duryea JD, Kessler NJ, Hellenthal G, Wood AC, Hodges KR, et al: **A Genomic Atlas of Systemic Interindividual Epigenetic Variation in Humans.** *Environmental and Molecular Mutagenesis* 2019, **60:**51-52.
27. Van Baak TE, Coarfa C, Dugue PA, Fiorito G, Laritsky E, Baker MS, Kessler NJ, Dong JR, Duryea JD, Silver MJ, et al: **Epigenetic supersimilarity of monozygotic twin pairs.** *Genome Biology* 2018, **19**.
28. McClay JL, Shabalin AA, Dozmorov MG, Adkins DE, Kumar G, Nerella S, Clark SL, Bergen SE, Hultman CM, Magnusson PKE, et al: **High density methylation QTL analysis in human blood via next-generation sequencing of the methylated genomic DNA fraction.** *Genome Biology* 2015, **16**.
29. Urdinguio RG, Bayon GF, Dmitrijeva M, Torano EG, Bravo C, Fraga MF, Bassas L, Larriba S, Fernandez AF: **Aberrant DNA methylation patterns of spermatozoa in men with unexplained infertility.** *Human Reproduction* 2015, **30:**1014-1028.
30. Rakyan VK, Down TA, Thorne NP, Flicek P, Kulesha E, Graf S, Tomazou EM, Backdahl L, Johnson N, Herberth M, et al: **An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs).** *Genome Research* 2008, **18:**1518-1529.
31. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, Van Djik S, Muhlhausler B, Stirzaker C, Clark SJ: **Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling.** *Genome Biology* 2016, **17**.
32. Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, Kim SY, Keefe DL, Alukal JP, Boeke JD, Yanai I: **Widespread Transcriptional Scanning in the Testis Modulates Gene Evolution Rates.** *Cell* 2020, **180:**248-262.e221.
33. Braun PR, Han SZ, Hing B, Nagahama Y, Gaul LN, Heinzman JT, Grossbach AJ, Close L, Dlouhy BJ, Howardiii MA, et al: **Genome-wide DNA methylation comparison between live human brain and peripheral tissues within individuals.** *Translational Psychiatry* 2019, **9**.
34. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R: **Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.** *Epigenetics* 2013, **8:**203-209.
35. Price EM, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS: **Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array.** *Epigenetics & Chromatin* 2013, **6**.
36. Kirchner H, Sinha I, Gao H, Ruby MA, Schonke M, Lindvall JM, Barres R, Krook A, Naslund E, Dahlman-Wright K, Zierath JR: **Altered DNA methylation of glycolytic and lipogenic genes in liver from obese and type 2 diabetic patients.** *Molecular Metabolism* 2016, **5:**171-183.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

37. Spiers H, Hannon E, Schalkwyk LC, Smith R, Wong CCY, O'Donovan MC, Bray NJ, Mill J: **Methylomic trajectories across human fetal brain development.** *Genome Research* 2015, **25:**338-352.
38. Illumina: **Pub. No. 1070-2015-008-B. Infinium MethylationEPIC BeadChip Datasheet.** Illumina; 2017.
39. Donkin I, Versteyhe S, Ingerslev LR, Qian K, Mechta M, Nordkap L, Mortensen B, Appel EVR, Jorgensen N, Kristiansen VB, et al: **Obesity and Bariatric Surgery Drive Epigenetic Variation of Spermatozoa in Humans.** *Cell Metabolism* 2016, **23:**369-378.
40. Camprubi C, Salas-Huetos A, Aiese-Cigliano R, Godo A, Pons MC, Castellano G, Grossmann M, Sanseverino W, Martin-Subero JI, Garrido N, Blanco J: **Spermatozoa from infertile patients exhibit differences of DNA methylation associated with spermatogenesis-related processes: an array-based analysis.** *Reproductive Biomedicine Online* 2016, **33:**709-719.
41. Jenkins TG, Aston KI, Meyer TD, Hotaling JM, Shamsi MB, Johnstone EB, Cox KJ, Stanford JB, Porucznik CA, Carrell DT: **Decreased fecundity and sperm DNA methylation patterns.** *Fertility and Sterility* 2016, **105:**51-+.
42. Gorrie-Stone TJ, Smart MC, Saffari A, Malki K, Hannon E, Burrage J, Mill J, Kumari M, Schalkwyk LC: **Bigmelon: tools for analysing large DNA methylation datasets.** *Bioinformatics* 2019, **35:**981-986.
43. World Health Organization: *WHO laboratory manual for the examination and processing of human semen- Fifth Edition.* Geneva, Switzerland: WHO; 2010.
44. Laqqan M, Tierling S, Alkhaled Y, LoPorto C, Hammadeh ME: **Alterations in sperm DNA methylation patterns of oligospermic males.** *Reproductive Biology* 2017, **17:**396-400.
45. Qiagen: **QIAamp. DNA Mini and Blood Mini Handbook 1102728.** vol. 1102728, Fifth edition edition: Qiagen HB-0329-004; May 2016.
46. Danson AF, Marzi SJ, Lowe R, Holland ML, Rakyan VK: **Early life diet conditions the molecular response to post-weaning protein restriction in the mouse.** *Bmc Biology* 2018, **16**.
47. Illumina: **Infinium HD Assay Methylation Protocol Guide Document # 15019519.** Illumina, Inc; 2015.
48. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC: **A data-driven approach to preprocessing Illumina 450K methylation array data.** *Bmc Genomics* 2013, **14**.
49. Ritchie ME, Phipson B, Wu D, Hu YF, Law CW, Shi W, Smyth GK: **limma powers differential expression analyses for RNA-sequencing and microarray studies.** *Nucleic Acids Research* 2015, **43**.
50. Filzmoser P, Maronna R, Werner M: **Outlier identification in high dimensions.** *Computational Statistics & Data Analysis* 2008, **52:**1694-1711.
51. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for Computing and Annotating Genomic Ranges.** *Plos Computational Biology* 2013, **9**.
52. Phipson B, Maksimovic J, Oshlack A: **missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform.** *Bioinformatics* 2016, **32:**286-288.
53. **Millard** SP: **EnvStats: An R Package for Environmental Statistics.** (Springer ed.: Springer; 2013.
54. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K: **cluster: Cluster Analysis Basics and Extensions. R package version 2.1.0.**; 2019.
55. Bates D, Machler M, Bolker BM, Walker SC: **Fitting Linear Mixed-Effects Models Using lme4.** *Journal of Statistical Software* 2015, **67:**1-48.
56. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT: **DNA methylation arrays as surrogate measures of cell mixture distribution.** *Bmc Bioinformatics* 2012, **13**.
57. Mansell G, Gorrie-Stone TJ, Bao YC, Kumari M, Schalkwyk LS, Mill J, Hannon E: **Guidance for DNA methylation studies: statistical insights from the Illumina EPIC array.** *Bmc Genomics* 2019, **20**.
58. Hillman S, Peebles DM, Williams DJ: **Paternal metabolic and cardiovascular risk factors for fetal growth restriction: a case-control study.** *Diabetes Care* 2013, **36:**1675-1680.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Material

Replication

The majority of DNA methylation differences observed between whole blood and sperm replicated in the lean replication (n = 21 lean males) and obesity cohort (n = 22 obese males) across the 692,219 probes that survived quality control in these cohorts: 288,062 of significant sites that were also present in the replication cohorts showed significant differences between blood and sperm in the replication cohort (65%; $P < 9 \times 10^{-8}$, paired t-test), and 306,023 sites (69%) in the obesity cohort. The effect sizes at the 441,764 significant probes from discovery, which were also present in the replication cohorts, were highly correlated with those observed in the replication cohorts (lean cohort: $r = 98\%$, $P < 1.0 \times 10^{-50}$; obese cohort: $r = 0.99$, $P < 1.0 \times 10^{-50}$; **Figure S5, Table S3**).

Correlations between whole blood and sperm DNA methylation were replicated in the two replication cohorts. 1,250 of the 1,513 significantly correlated sites had also passed quality control in the replication cohorts and 455 (36%) of these were significantly correlated in the lean replication cohort ($P < 9 \times 10^{-8}$, Pearson's product moment correlation), 502 (40%) in the obesity cohort (**Table S12**). Given the reduced power to detect significant correlation in these two cohorts of reduced size, we further characterized sites showing very little evidence of correlation in the replication of cohorts ($r < 0.3$ in both cohorts). These 173 sites (14%) are all driven by groups of outliers in the discovery cohort resulting in bi- or tri-modal distribution in the discovery sample, that were not present in the replication cohorts (examples shown in **Figure S7**). The majority (127 sites; 73%) were characterized by a bimodal distribution with a single outlier in the discovery.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Tables

CpG region	DNA methylation	P	OR
Island	High	3.90E-15	0.85
Island	Low	< 1.00E-50	35.60
Shore	High	< 1.00E-50	0.27
Shore	Low	< 1.00E-50	1.95
Shelf	High	0.11	1.02
Shelf	Low	< 1.00E-50	0.11
Open sea	High	< 1.00E-50	2.19
Open sea	Low	< 1.00E-50	0.07

Supplementary Table 1. Enrichments of CpG region annotations across sites showing extreme methylation values in sperm.
Sites showing > 0.8 median beta value were classified as “high”, sites with median beta < 0.2 as “low”. Enrichments of each region amongst “high” and “low” methylation sites were calculated against the annotation of intermediately methylated sites (median beta between 0.2 and 0.8) using a Fisher’s exact test.
OR = odds ratio

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Region	DNA methylation	P	OR
TSS1500	High	< 1.00E-50	0.41
TSS1500	Low	< 1.00E-50	2.12
TSS200	High	< 1.00E-50	0.66
TSS200	Low	< 1.00E-50	6.59
5'UTR	High	2.21E-16	0.91
5'UTR	Low	< 1.00E-50	1.78
Body	High	< 1.00E-50	1.59
Body	Low	< 1.00E-50	0.40
1st exon	High	< 1.00E-50	0.67
1st exon	Low	< 1.00E-50	6.17
Exon boundary	High	5.39E-46	1.62
Exon boundary	Low	< 1.00E-50	0.22
3'UTR	High	3.81E-10	1.13
3'UTR	Low	< 1.00E-50	0.36
Not annotated	High	< 1.00E-50	0.87
Not annotated	Low	< 1.00E-50	0.31

Supplementary Table 2. Enrichments of genomic region annotations across sites showing extreme methylation values in sperm.
Sites showing > 80% median DNA methylation were classified as “high”, sites with < 20% methylation as “low”. Enrichments of each region amongst “high” and “low” methylation sites were calculated against the annotation of intermediately methylated sites (20-80% median DNA methylation) using a Fisher’s exact test.
OR = odds ratio

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 3. Summary statistics for differences in DNA methylation between whole blood and sperm.
We used a paired t-test to identify DNA methylation differences between whole blood and sperm across all 704,356 probes passing quality control in the discovery dataset. Summary statistics are reported for all sites in the discovery dataset. Summary statistics from the replication cohort are reported for sites that also passed quality control in our replication dataset.
IlmnID = Illumina CpG identifier, chr = chromosome, location = position on chromosome in hg19 reference, P = p-value in the discovery data, effect = effect size in the discovery data, P_rep = p-value in the lean replication cohort, effect_lean = effect size in the lean replication cohort, P_ob = p-value in the obese replication cohort, effect_ob = effect size in the obese replication cohort.

Table S3 is part of the electronic appendix:
SupplementaryTable3_BloodSpermDiff.csv

	Discovery cohort	Lean replication cohort	Obesity cohort	P (difference between cohorts)	P (discovery vs replication)	P (discovery vs obesity)	P (replication vs obesity)
Age (years). Mean (SD)	36.3 (5.2)	34.1 (4.6)	35.1 (4.1)	0.192			
BMI (kg/m2). Mean (SD)	23.4 (4.6)	22.3 (1.1)	29.1 (3.2)	<0.001	0.060	<0.001	<0.001
Waist circumference (cm). Mean (SD)	84.4 (4.8)	82.4 (6.4)	99.4 (8.7)	<0.001	0.436	<0.001	<0.001
SPB (mmHg), average of two measurements. Mean (SD)	119 (11)	121 (10)	126 (9)	0.052			
DPB (mmHg), average of two measurements. Mean (SD)	77 (8)	78 (6)	81 (8)	0.050			
Total cholesterol (mmol/L). Mean (SD)	4.7 (0.7)	4.9 (0.9)	4.9 (1)	0.614			
HDL cholesterol (mmol/L). Mean (SD)	1.6 (0.3)	1.5 (0.3)	1.4 (0.3)	0.060			
LDL cholesterol (mmol/L). Mean (SD)	2.7 (0.7)	2.9 (0.8)	2.9 (0.9)	0.330			
Fasting glucose (mmol/L). Median (IQR)	4.8 (0.5)	4.6 (0.4)	4.7 (0.6)	0.018	0.003	0.088	0.105
Fasting insulin (mIU/L). Median (IQR)	5.3 (3.4)	5.1 (3.0)	8.9 (7.2)	0.002	0.309	<0.001	0.004
HOMA-IR. Median (IQR)	1.2 (0.8)	1.1 (0.6)	1.9 (1.4)	<0.001	0.285	<0.001	0.005
HOMA2-IR. Median (IQR)	1.1 (0.5)	0.6 (0.4)	1.1 (0.9)	0.014	0.048	0.414	0.003
CRP (mg/L). Median (IQR)	0.6 (0.3)	0.6 (0.1)	1 (1.8)	<0.001	0.105	0.001	<0.001
Triglycerides (mmol/L). Median (IQR)	0.9 (0.5)	0.9 (0.7)	1.2 (0.6)	0.282	0.335	0.056	0.157

Supplementary Table 4. Phenotype characteristics of participants included in the discovery, replication and obesity cohorts.
Reference ranges are derived from the UCLH Clinical Biochemistry Test Information sheet available from (1). The reference range for HOMA-IR is derived from (2). SD = Standard Deviation, IQR = interquartile range, BMI = Body Mass Index, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, HOMA-IR = Homeostatic Model Assessment of Insulin Resistance, CRP = C-Reactive Protein, HDL = High Density Lipoprotein, LDL = Low Density Lipoprotein

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CpG region	Probes	P	DNA methylation difference (beta)
Island	132,883	< 1.00E-50	-0.07
Shore	128,079	< 1.00E-50	-0.16
Shelf	48,301	< 1.00E-50	0.06
Sea	395,093	< 1.00E-50	0.07

Supplementary Table 5. Blood and sperm DNA methylation difference by CpG region.
Using a paired t-test the DNA methylation difference between the median methylation in blood and sperm was calculated for each region. The DNA methylation difference is shown with respect to blood (a positive value indicating higher average DNA methylation in sperm).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Region	Probes	P	DNA methylation difference (beta)
TSS1500	103,486	< 1.00E-50	-0.11
TSS200	64,958	< 1.00E-50	-0.02
5'UTR	92,296	3.61E-32	0.09
Body	297,434	< 1.00E-50	0.02
1st exon	38,767	< 1.00E-50	-0.02
Exon boundary	6,462	2.80E-22	0.02
3'UTR	20,248	3.81E-05	-0.01
Not annotated	191,155	< 1.00E-50	0.02

Supplementary Table 6. Blood and sperm DNA methylation difference by genomic region.
Using a paired t-test the DNA methylation difference between the median methylation in blood and sperm was calculated for each region. The DNA methylation difference is shown with respect to blood (a positive value indicating higher average DNA methylation in sperm).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Tissue	Number of samples
Adipose	42
Blood	2317
Brain	868
Buccal	214
Cartilage	60
Chorion	3
Colon	170
Epithelial	183
Fibroblast	54
Intestines	1
Kidney	45
Liver	90
Lung	103
Lymph node	24
Mucosa	95
Muscle	17
Neuron	71
Neutrophils	69
Pancreas	112
Rectum	13
Saliva	146
Skin	38
T cells	136
Unsorted cell lines	9
Unsorted tissues	863
Unsorted tumours	174

Supplementary Table 7. Details on non-sperm tissue samples in the GEO analysis. The corresponding accession numbers are provided in Table S8.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 8. Accession numbers of all DNA methylation samples downloaded from GEO.

Table S8 is part of the electronic appendix:
SupplementaryTable8_GEOAccessionNumbers.csv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 9. Summary statistics for differences in DNA methylation between sperm and somatic tissue samples from GEO.

We compared DNA methylation in 371 sperm samples (including 90 samples from our cohorts) to that of 5,917 somatic tissue samples from GEO using linear regression. This analysis was conducted across all 452,626 sites that are present on both the 450K and EPIC array. Summary statistics are reported for all sites.

IlmnID = Illumina CpG identifier, chr = chromosome, location = position on chromosome in hg19 reference, P = P value for difference between sperm and somatic cell DNA methylation, P_Bonferroni = Bonferroni-adjusted P value, effect = DNA methylation difference (beta) – negative values indicate lower DNA methylation in sperm compared to somatic tissues.

Table S9 is part of the electronic appendix:

SupplementaryTable9_GEOSummaryStatistics.csv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 10. Significantly enriched Gene ontology terms amongst CpG sites identified to be hypermethylated in sperm compared to somatic tissues.

GO analysis identified 272 pathways enriched amongst hypermethylated sites. Of note, 37 of these (14%) related to transcriptional regulation, while 67 (25%) were related to brain and neurological categories.

GO ID = Gene Ontology identifier, N = number of genes in the GO term, DE = number of genes that were differentially methylated, P.DE = P value for over-representation of the GO term, ONTOLOGY: BP = biological process, CC = cellular component, MF = molecular function

Table S10 is part of the electronic appendix:
SupplementaryTable10_HyperPathways.csv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 11. Significantly enriched Gene ontology terms amongst CpG sites identified to be hypomethylated in sperm compared to somatic tissues.

GO analysis identified 37 pathways enriched amongst hypomethylated sites. Eight of these pathways were related to sensory perception, specifically smell.

GO ID = Gene Ontology identifier, N = number of genes in the GO term, DE = number of genes that were differentially methylated, P.DE = P value for over-representation of the GO term, ONTOLOGY: BP = biological process, CC = cellular component, MF = molecular function

Table S11 is part of the electronic appendix:
SupplementaryTable11_HypoPathways.csv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 12. Summary statistics for correlation of DNA methylation between whole blood and sperm.
We used a Pearson's correlation test to identify CpG sites where DNA methylation was significantly correlated between whole blood and sperm This analysis was restricted to the 155,269 sites that showed met minimum variability criteria in both tissues (range of middle 80% > 5%). Summary statistics are reported for all sites in the discovery dataset. Summary statistics from the replication cohort are reported for the sites that also passed quality control in our replication dataset.
IlmnID = Illumina CpG identifier, chr = chromosome, location = position on chromosome in hg19 reference, P = p-value in the discovery data, r = correlation coefficient in the discovery data, P_rep = p-value in the lean replication cohort, r_lean = correlation coefficient in the lean replication cohort, P_ob = p-value in the obese replication cohort, r_ob = correlation coefficient in the obese replication cohort.

Table S12 is part of the electronic appendix:
SupplementaryTable12_Correlations.csv

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Table 13. Summary statistics for the association between DNA methylation and obesity in whole blood and sperm.

We regressed DNA methylation onto obesity status in our replication cohort, separately in whole blood and sperm, controlling for estimated blood cell type proportions in the blood analysis. We furthermore used a linear mixed effects model across the combined discovery and replication datasets, regressing DNA methylation onto obesity status, tissue type and batch while controlling for interindividual variation. Summary statistics for both analyses are reported – the LME results are restricted to sites available in both the discovery and replication datasets.

IlmnID = Illumina CpG identifier, chr = chromosome, location = position on chromosome in hg19 reference, P_blood = p-value in blood analysis, effect_blood = effect size in whole blood, P_sperm = p-value in sperm analysis, effect_sperm = effect size in sperm, P_mix = p-value in the mixed effects model, effect_mix= effect size in the mixed effects model. All effect sized are reported using the lean men as reference group.

Table S13 is part of the electronic appendix:
SupplementaryTable13_ObesityAssociations.csv

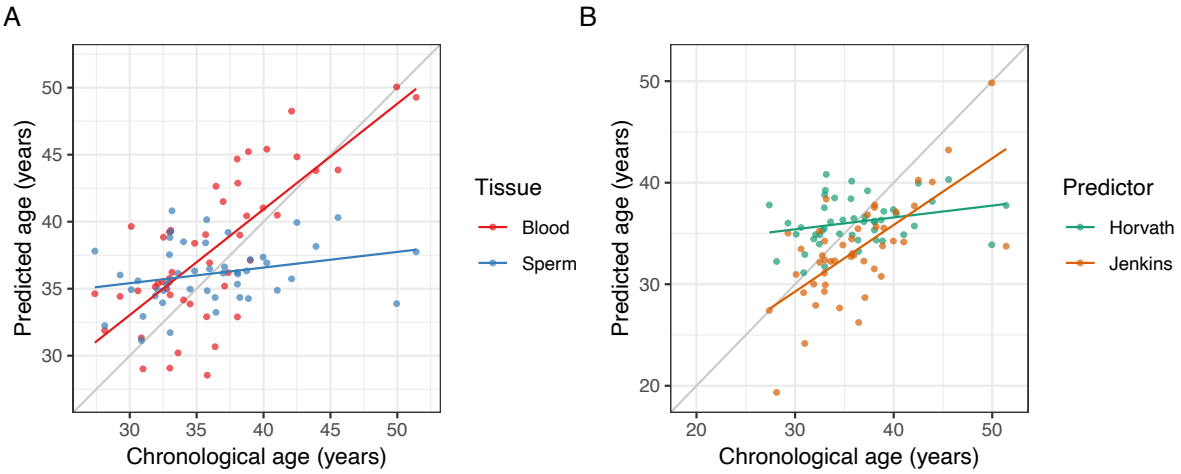
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

	Discovery cohort	Lean replication cohort	Obesity cohort	Reference range	P
Volume (sperm, mL). Mean (SD)	2.9 (1.1)	2.9 (1.4)	2.6 (1.5)	> 1.5 mL	0.538
Concentration (sperm, millions). Mean (SD)	55.4 (37.2)	47.9 (33.9)	57.4 (31)	> 15 millions/mL	0.608
Total count per ejaculate (millions). Mean (SD)	161 (150.4)	149 (140.5)	157 (131.5)	> 39 million	0.953
Percentage A sperm. Mean (SD)	14.8 (10.6)	15.4 (10.6)	17.4 (10.3)	N/A	0.610
Percentage B sperm. Mean (SD)	23.9 (9.5)	22.1 (9.4)	20.4 (8.6)	N/A	0.348
Percentage C sperm. Mean (SD)	12.1 (3.7)	11.4 (3.7)	11.1 (4.3)	N/A	0.589
Percentage D sperm. Mean (SD)	49.3 (18.1)	50.5 (18.7)	51.1 (18.8)	N/A	0.926
Average motile speed. Mean (SD)	18.6 (2.6)	19.2 (4.4)	19.4 (2.3)	N/A	0.603

Supplementary Table 14. Semen sample parameters for the discovery and replication cohorts (the lean replication cohort and the obesity cohort). Semen sample parameters were measured using the Computer-Assisted Sperm Analysis (CASA)/Sperminator software (Pro-Creative Diagnostics, Staffordshire, UK). V = volume, C = concentration, SD = Standard Deviation, WHO = World Health Organization. Percentage A-D sperm refer to the proportion of spermatozoa in different motility grades where A = most motile and D = least motile. Reference ranges derived from (3).

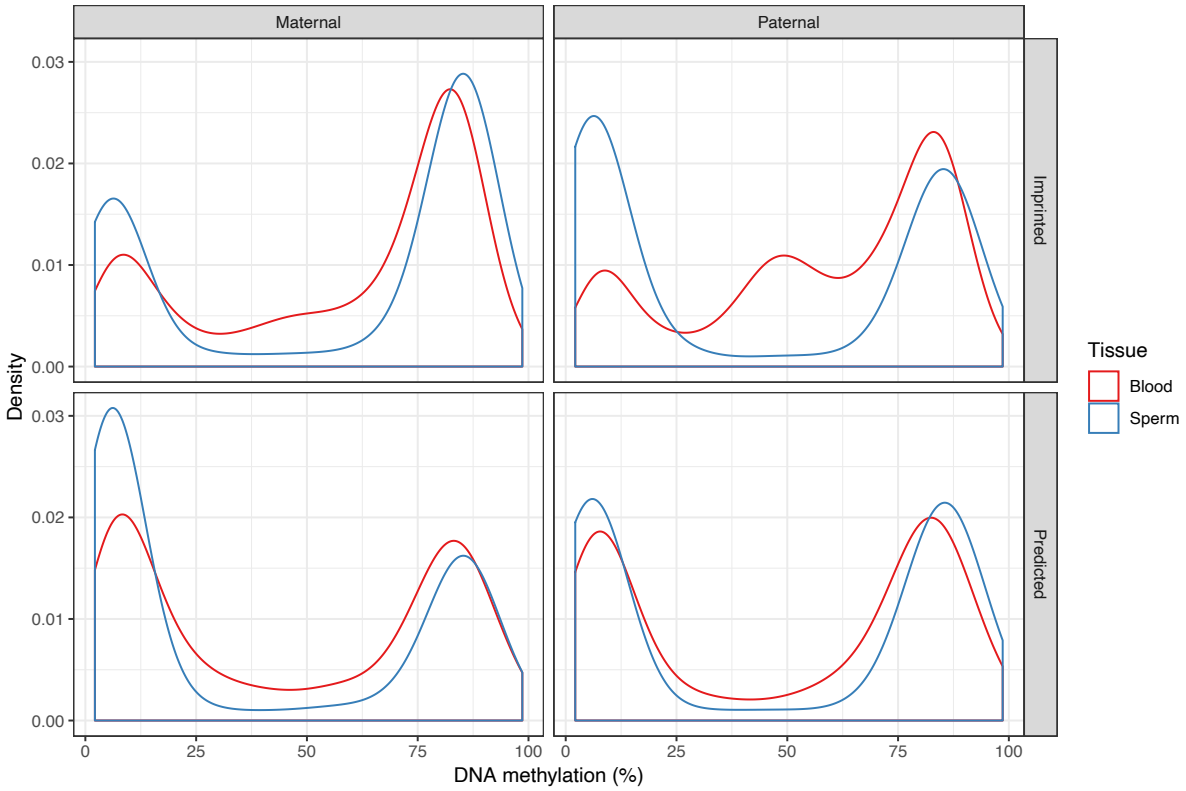
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figures



Supplementary Figure 1. DNA methylation age prediction in whole blood and sperm.
A) As reported previously, the DNA methylation age predictor by Horvath was able to accurately predict chronological age from DNA methylation in whole blood ($r = 0.74$, $P = 2.55 \times 10^{-9}$, Pearson's product moment correlation) but not in sperm ($r = 0.26$, $P = 0.07$).
B) However, chronological age could be more accurately predicted from DNA methylation in sperm using the predictor more recently developed by Jenkins and colleagues ($r = 0.68$, $P = 1.78 \times 10^{-7}$).

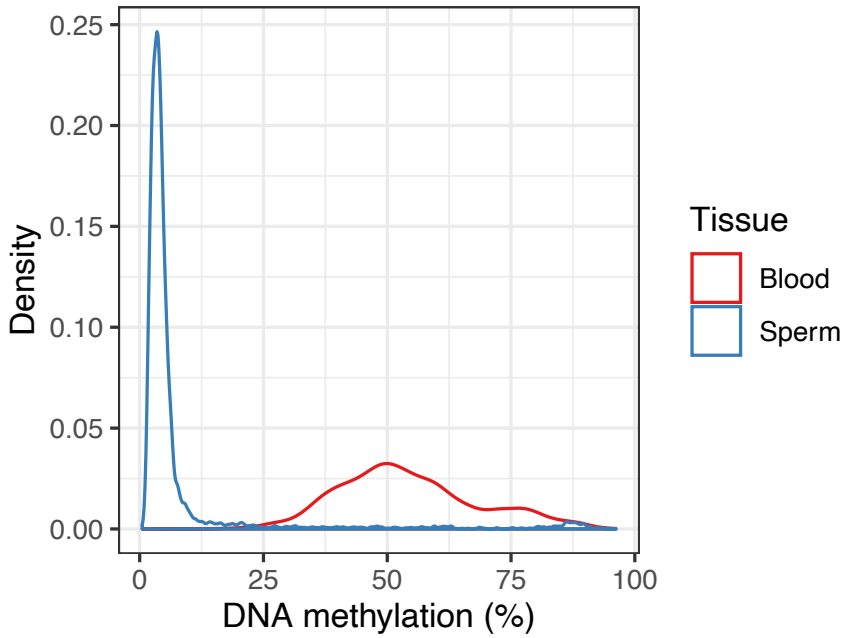
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 2. DNA methylation at CpG sites annotated to imprinted genes is enriched in intermediate levels of DNA methylation in blood, but not sperm.

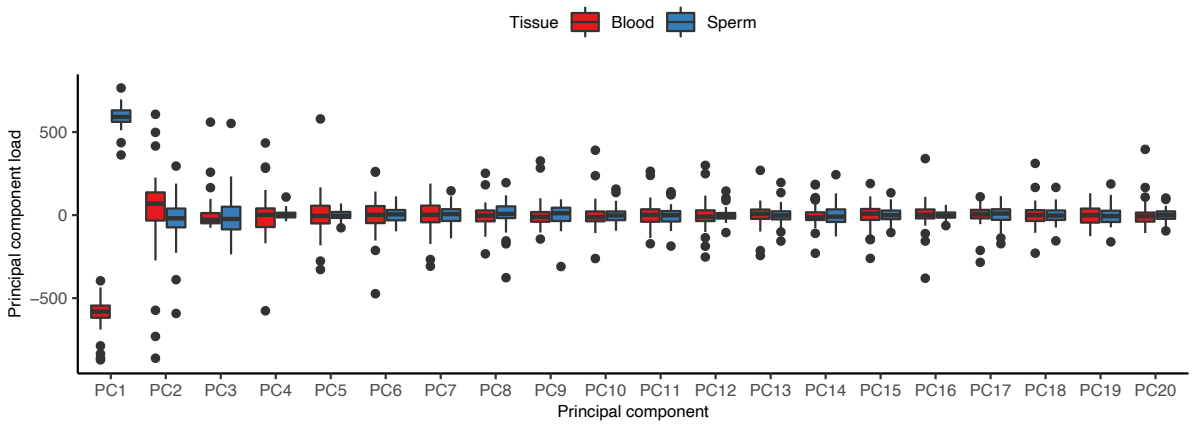
DNA methylation annotated to known imprinted genes (Geneimprint database; http://www.geneimprint.com), showed a characteristic enrichment in sites with beta around 0.5 (+/- 0.1) in whole blood – particularly, those genes known to be paternally imprinted ($P < 1.00 \times 10^{-50}$, Fisher's exact test), but also for maternally imprinted genes ($P = 9.19 \times 10^{-9}$) and a less pronounced enrichment in genes predicted to be imprinted paternally ($P = 0.01$) or maternally ($P = 0.04$). No such enrichment was observed in sperm ($P > 0.05$ for all four tests).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



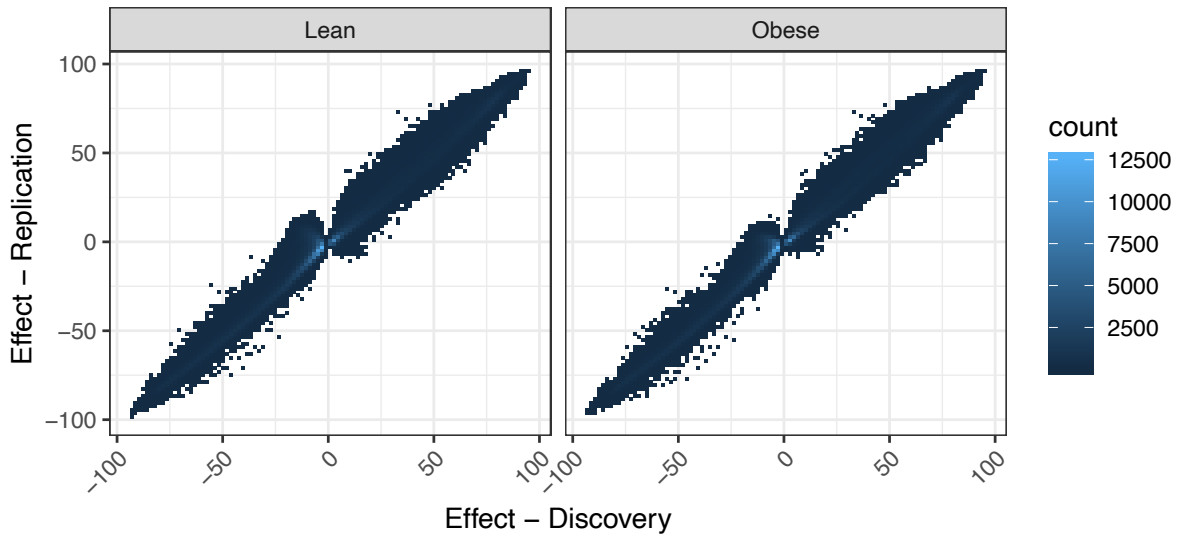
Supplementary Figure 3. Imprinting control regions are around 50% methylated in whole blood, while being nearly completely unmethylated in sperm.
Nearly all of the 169 CpG sites that are located in known imprinting control regions (ICRs, WAMIDEX database; https://atlas.genetics.kcl.ac.uk) display intermediate DNA methylation levels in blood (57% of sites with median beta between 0.4 and 0.6; $P < 1.00 \times 10^{-50}$, Fisher's exact test). Simultaneously, they appear to be completely unmethylated in sperm (94% of sites with median beta < 0.2, $P < 1.00 \times 10^{-50}$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 4. Load of DNA methylation on first 20 principal components (PCs) in whole blood and sperm. The first PC, which explained 51.41% of the total variance, clearly distinguishes between blood and sperm, making tissue/cell type the single biggest factor contributing to variation in DNA methylation across our samples.

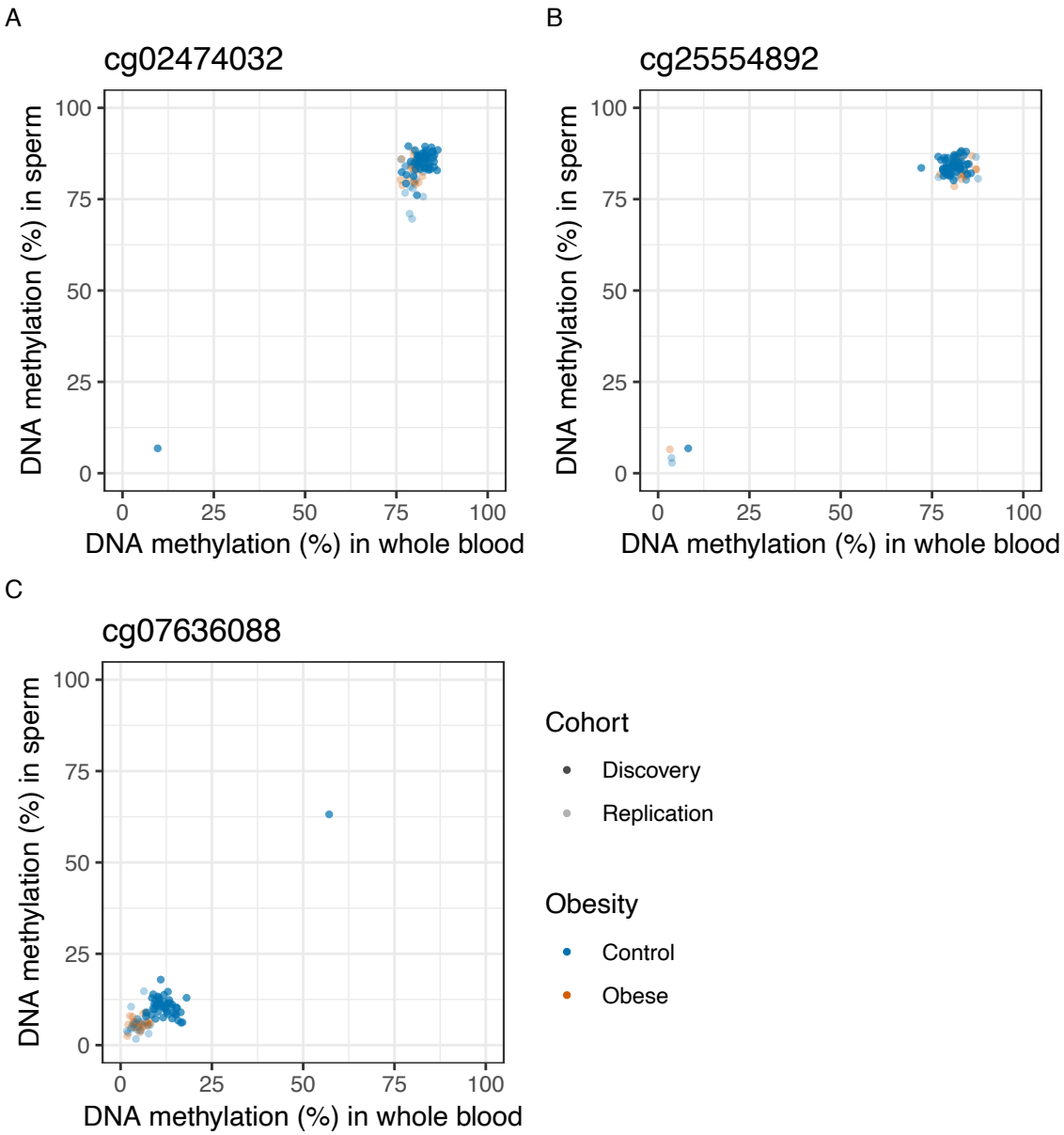
bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 5. Differences observed between whole blood and sperm DNA methylation replicated across two replication cohorts.

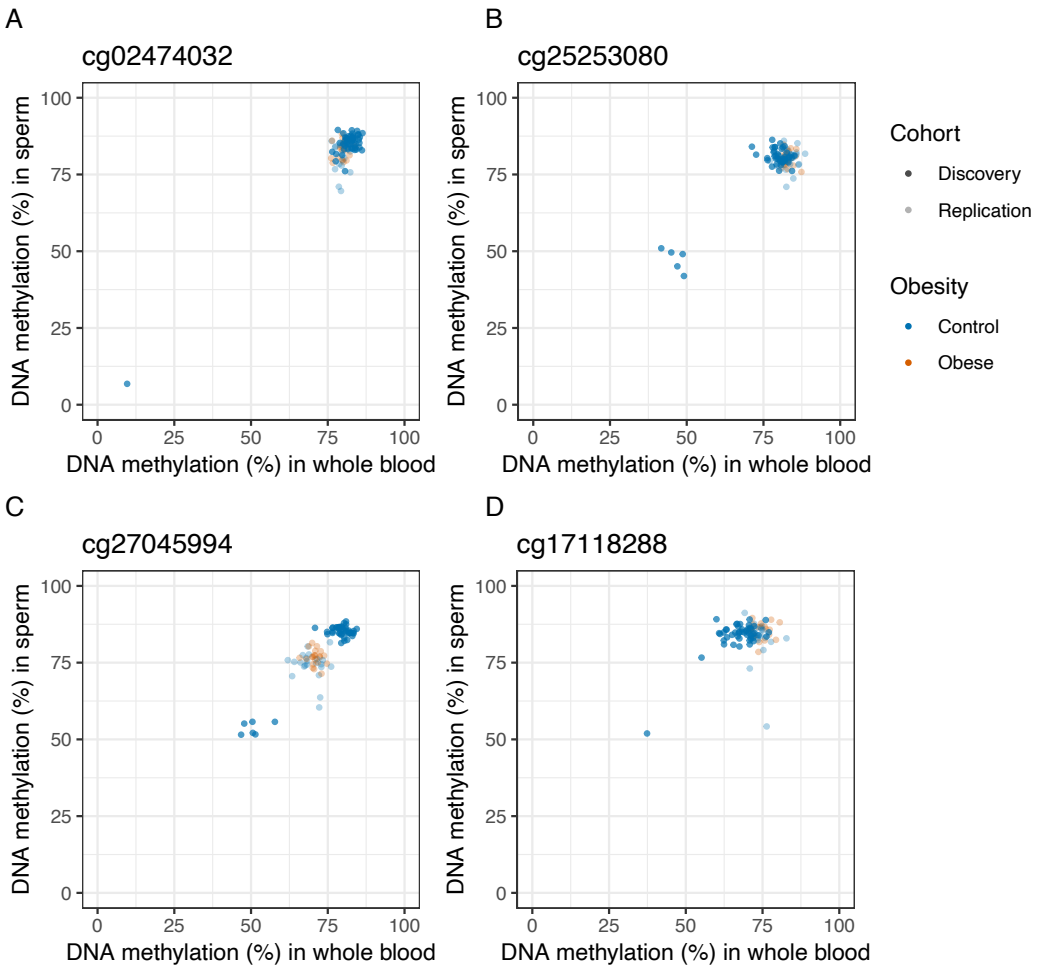
The effect sizes at the 441,764 significant probes from discovery, which were also present in the replication cohorts, were highly correlated with those observed in the replication cohorts (lean cohort: $r = 98\%$, $P < 1.0 \times 10^{-50}$; obese cohort: $r = 0.99$, $P < 1.0 \times 10^{-50}$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 6. 365 of the 1,513 significantly correlated sites were driven by single outliers.
Shown is DNA methylation in whole blood and sperm from the discovery and replication cohorts at
A) cg02474032 (chr16:87678659),
B) cg25554892 (chrX:70434406), and
C) cg07636088 (chr13: 31734946). We observed higher measured DNA methylation in the individual outlier at less than 2% of these 365 sites.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 7. Correlations which did not replicate were driven small numbers of individual outliers in the discovery cohort.

Of the 1,250 correlated probes also present in the replication cohorts 173 (13%) show no evidence of correlation in the replication cohorts ($r < 0.3$ in both cohorts)

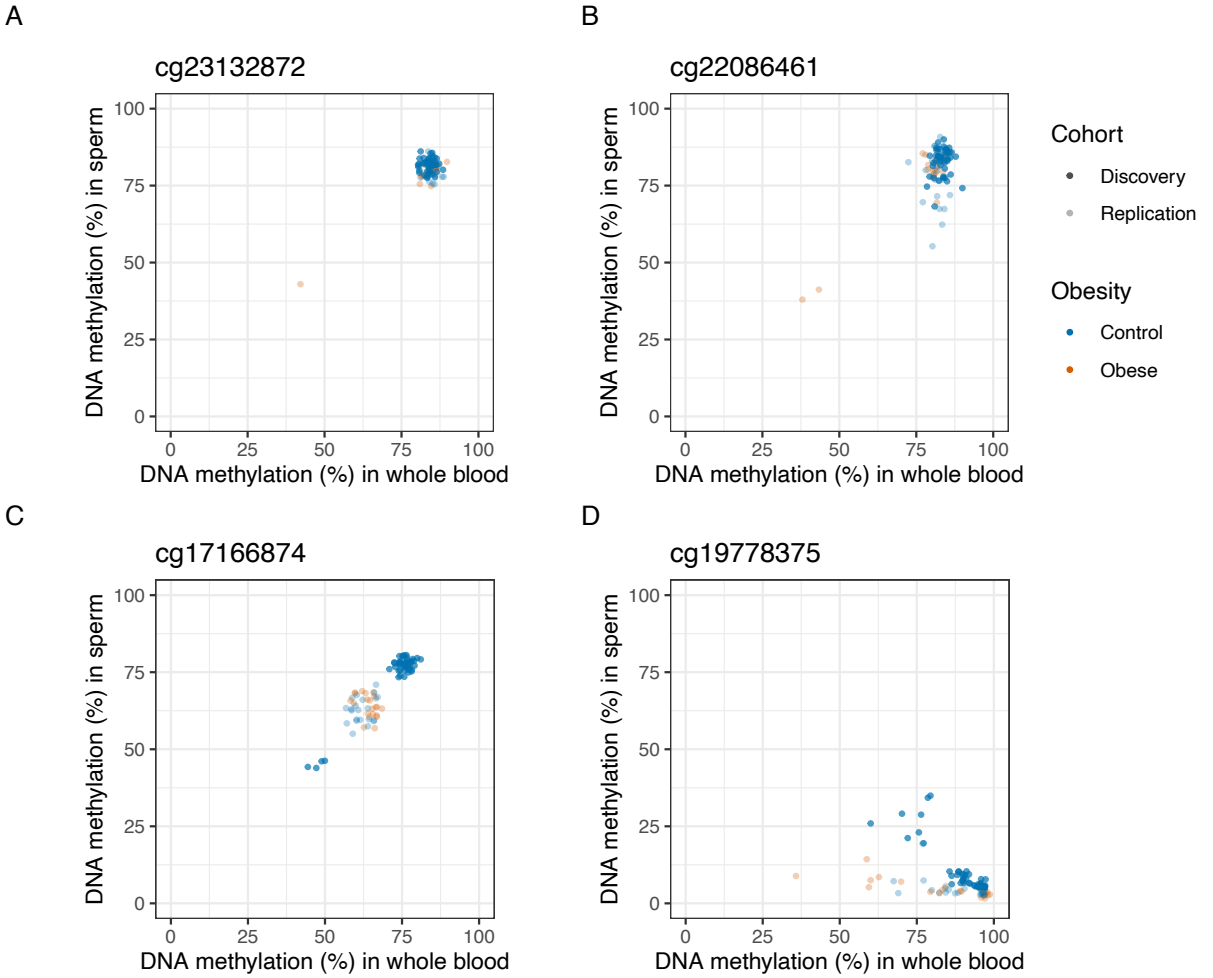
A) The majority of these sites (127 sites; 76%) were characterized by a single outlier in the discovery cohort, without any outliers in the replication cohorts. One example is found at cg27045994 (chr16:87678659).

B) cg25253080 (chr10:14795564) represents the only incidence where a group of 5 outliers did not replicate in either replication cohort.

C) The biggest outlier group which did not replicate contained 6 individuals, with no outliers in the replication data and was found at cg27045994 (chr8:284126).

D) The only trimodal distribution which did not replicate was observed at cg17118288 (chr1:218563763).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 8. Statistically significant interaction effects were driven by outliers in either the obese or lean group.

The majority of significant interactions between sperm and blood DNA methylation and obesity were driven by single or very few outliers in the obesity group.

A) At cg23132872 (chr2:191882300), the correlation in obese individuals is driven by a single outlier.

B) At cg22086461 (chr8:77343728) the correlation in obese individuals is driven by two outliers.

C) At cg17166874 (chr7:155381422) the correlation in lean men is driven by four outliers in the discovery cohort and methylation at this site is also characterized by substantial batch effects.

D) At cg19778375 (chr12:297831) there appears to be a batch effect between the discovery and replication cohort that contributes to an observed correlation in the lean men from the discovery cohort, which is not present in the replication datasets.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.01.072934; this version posted May 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary References

1. UCLH Clinical Biochemistry. UCLH Clinical Biochemistry Test Information University College London Hospital2017 [Biochemistry test information]. Available from: https://www.uclh.nhs.uk/OurServices/ServiceA-Z/PATH/PATHBIOMED/CBIO/Pages/InformationforGPs.aspx.
2. Gayoso-Diz P, Otero-Gonzalez A, Rodriguez-Alvarez MX, Gude F, Garcia F, De Francisco A, et al. Insulin resistance (HOMA-IR) cut-off values and the metabolic syndrome in a general adult population: effect of gender and age: EPIRCE cross-sectional study. Bmc Endocrine Disorders. 2013;13.
3. World Health Organization. WHO laboratory manual for the examination and processing of human semen- Fifth Edition. WHO, editor. Geneva, Switzerland: WHO; 2010.