Supplementary Online Content

Aberg KA, McClay JL, Nerella S, et al. Methylome-wide association study of schizophrenia: indentifying blood biomarker signatures of environmental insults. *JAMA Psychiatry*. Published January 8, 2014. doi:10.1001/jamapsychiatry.2013.3730.

This supplementary material has been provided by the authors to give readers additional information about their work.

SUPPLEMENTAL MATERIAL FOR THE PAPER:

Methylome-wide sequencing study of schizophrenia identifies blood biomarker signatures of environmental insults

MBD-SEQ LABORATORY METHODS

In this section, we briefly describe the laboratory methods. A table with descriptive statistics is provided at the end of the section. In the text below, **bold** indicates that the entry will appear in this table.

DNA extraction DNA samples from all participants were extracted using the buffy coat from EDTA blood using the Gentra Puregene kit for automated extraction with the Autopure LS robot following the manufacturer's instruction (Qiagen, Valencia, CA). Prior to methylation investigations, the DNA quality was evaluated using 1% agarose gels. Concentration and 260/280 ratios were measured with Nanodrop1000 (Thermofisher, Waltham, MA). Samples were then shipped to the Broad Institute at Harvard/MIT where they were fingerprinted using a 24-plex sequenom fingerprint iPLEX assay prior to shipping to the service provider that did the sequencing for this project.

Fragmentation DNA was fragmented to an approximate median fragment size of 125bp, using ultrasonication with the Covaris E210 system (Covaris, Woburn, MA). The shearing protocol consisted of 10 cycles (duty cycle 10%, intensity 5, cycles/burst 100) in a bath temperature of 10°C. To confirm successful fragmentation, all samples were evaluated on an Agilent Bioanalyzer using the DNA1000 chip kit (Agilent, Santa Clara, CA). Post-fragmentation DNA concentration was measured with Nanodrop1000.

Enrichment protocol We used the MethylMiner DNA enrichment kit (Invitrogen, Carlsbad, CA), which employs the methyl-binding domain 2 (MBD2) protein to capture fragments with one or multiple methylated CpGs. Dependent on sample availability, 2-5µg of DNA (**Starting DNA**) was used as starting material for the methylation enrichment protocol. We followed the manufacturers' protocol for ≤5 µg of starting material and eluted the captured fragments using 500mM NaCl. The captured material was quantified using a Qubit fluorometer with Quant-It dsDNA HS assay kit (Invitrogen, Carlsbad, CA).

Library construction We constructed barcoded fragment libraries for SOLiD next-generation sequencing (Life Technologies, Foster City, CA) using the captured methylation-enriched DNA (**Mass captured**). Using barcodes 1-16 (Life Technologies, Foster City, CA) and following the manufacturer's protocol, the captured DNA was end-repaired and barcoded adaptors were ligated to each sample. To confirm successful adaptor ligation approximately 20% of the samples were evaluated on an Agilent Bioanalyzer using the Agilent high sensitivity DNA chip kit (Agilent, Santa Clara, CA). Next, all samples were quantified using the Quant-It and 6-8 samples with different barcodes were pooled in equal molarities. Following the manufacturer's instructions, the pools were nick-translated and amplified to obtain sufficient amounts of libraries. Completed libraries were quantified with Quant-It.

Emulsion PCR and next-generation sequencing Automated emulsion PCR (ePCR) was conducted using standard procedures for the SOLiD EZ bead system (Life Technologies, Foster City, CA). After ePCR, the beads with the barcoded fragment libraries were deposited to slides and sequenced with 50bp single-end chemistry, in addition to the length of the barcodes, on SOLiD3 plus and SOLiD4 instruments (Life Technologies, Foster City, CA).

Reads and Alignment Based on observations that 30–60 million reads per sample may be sufficient to reveal valuable information for whole-genome methylation analysis[1-2], we aimed for

more than 60 million reads per sample with a minimum of 40 million reads for each of the 1575 methylomes sequenced in this investigation. The 205 samples for which we got fewer than 40 million reads were rerun to supplement reads (**Rerun**). After these reruns, only 25 samples still had fewer than 40 million reads. The SOLiD system essentially scans each base twice thereby producing 2 "color calls" for each base. We deleted all reads with > 2 missing color calls (**Missing calls**) after which we observed an average of 68.0 million reads (**Total reads**) per sample.

The sequenced reads were aligned to the human genome (build hg19/GRCh37) using BioScope 1.2 (Life Technologies, Foster City, CA) that aligns in color-space and takes full advantage of the increased ability of SOLiD two-base encoding to identify sequencing errors [31]. We used a seed-and-extend approach combined with local alignment and multiple schemas. Specifically, our seed was 25 bases. Rather than considering the entire extension, local alignment may improve sensitivity by finding the maximum similarity score between the reference sequence and a substring of the extension (**Alignment length** includes the 25bp seed + the length of the extension). A maximum of two color space mismatches (**Sum mismatch** gives the total number of mismatches per sample) were allowed in the seed (e.g., as two color call matches are required to change the base call, a SNP will have two color call mismatches). If the seed could not be mapped, a second schema was attempted by moving the seed from base 1 to base 15. After dropping all reads from runs with <40% alignment or that produced a very small number of reads (these involved mainly reruns), the percentage of mapped reads was 69.8% (**Mapped reads**).

Many reads map to multiple locations of the genome. Often a single alignment can be selected because it is clearly better than the others (**SBA reads**, single best alignment). In the case of multi-reads, multiple alignments are approximately equivalent. Duplicate-reads (**Duplicate reads**) are reads that start at the same nucleotide positions. When sequencing a whole genome duplicate-reads often arise from template preparation or amplification artifacts. In our context of sequencing an enriched genomic fraction, duplicate-reads are increasingly likely to occur by chance because reads are expected to align to a much smaller fraction of the genome. We eliminated 32.5% of reads because they were (low quality) multi- or duplicate-reads (see above for definition). This left on average over 32 million reads/sample (**Used reads**).

Descriptive statistics eTable 1 provides descriptive statistics for the indices that are in **bold** in the section above.

eTable 1. Descriptive statistics for sequencing parameters

	All		Control		Case		Cohen's d	T	P
	Mean	SD	Mean	SD	Mean	SD			
Starting DNA (µg)	4.150	0.997	4.131	0.983	4.168	0.983	0.037	-0.713	0.476
Mass captured (µg)	0.176	0.079	0.174	0.081	0.177	0.081	0.038	-0.718	0.473
Rerun (yes/no)	0.085	0.279	0.091	0.271	0.080	0.271	0.039	0.746	0.456
Missing calls (mil)	87.593	35.329	86.944	34.869	88.221	34.869	0.036	-0.690	0.490
Total reads (mil)	68.021	26.830	67.336	26.719	68.684	26.719	0.050	-0.960	0.337
Alignment length (bp)	42.233	12.975	42.475	12.785	41.999	12.785	0.037	0.701	0.483
Sum mismatch (mil)	281.52	111.21	276.28	111.72	286.60	111.72	0.093	-1.774	0.076
Mapped reads (pro)	0.698	0.063	0.692	0.063	0.706	0.063	0.212	-4.062	0.000
SBA reads (pro)	0.765	0.029	0.763	0.030	0.767	0.030	0.108	-2.058	0.040
Duplicate reads (pro)	0.178	0.054	0.182	0.056	0.174	0.056	0.151	2.889	0.004
Used reads (mil)	32.443	13.685	31.669	13.934	33.192	13.934	0.111	-2.129	0.033

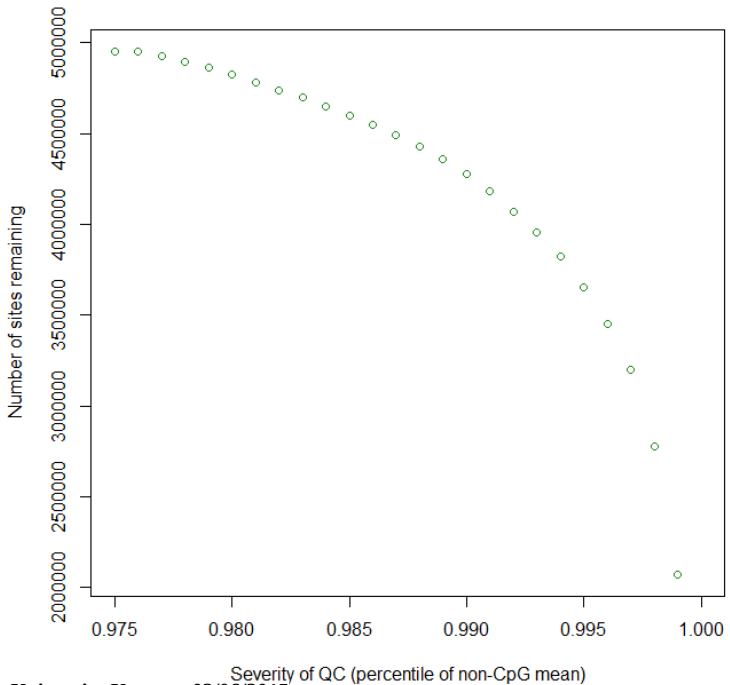
Mil = million, pro = proportion

Results show that these indices are generally equivalent for cases and controls. After a multiple testing correction, reads seemed to map slightly better in cases (69.2% vs. 70.6%), and also showed fewer duplicate reads (18.2% vs 17.4%). We note, however, that the differences are small and that all our coverage calculations are standardized on the total number of used reads.

1. Additional MWAS results and plots

eFigure 1. Severity of block QC versus number of sites remaining in the analysis

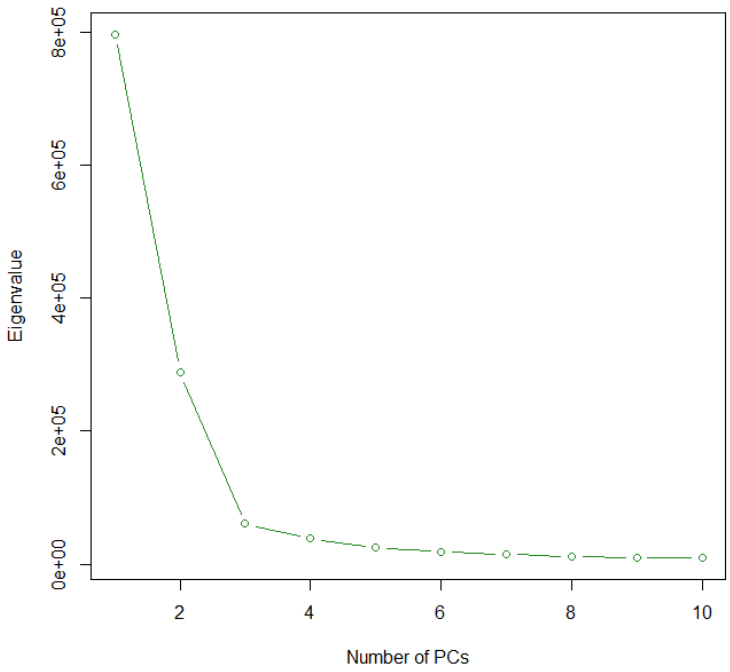
eFigure 1 shows the relationship between severity of block QC and the number of sites remaining in the analyses. e**Figure 3** shows that the number of sites that would be eliminated increases substantially after the 99th percentile of the coverage estimates at these non-CpGs. The 99th percentile means that only 1% of the CpG



sites that are not methylated (which is generally assumed to be about 20% of all CpG sites) will be included in the analysis. This threshold, therefore, strikes a good balance between the need of eliminating non-methylated sites versus retaining as complete as possible methylome-wide coverage.

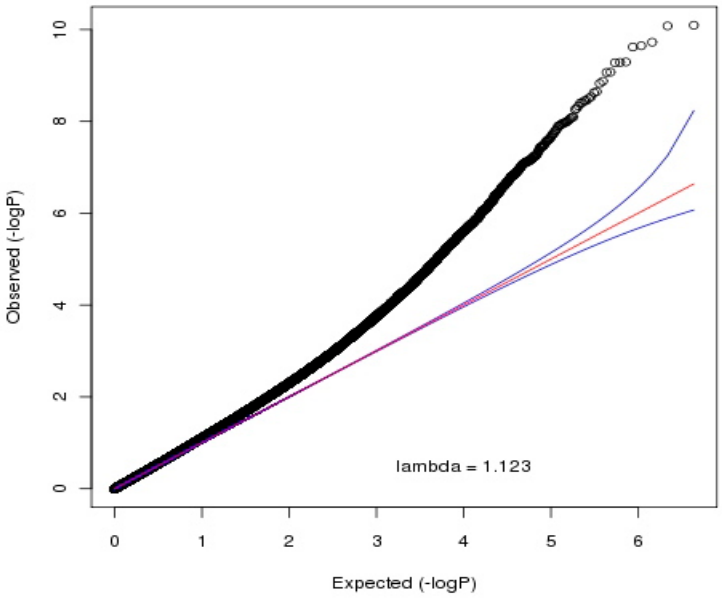
eFigure 2. Scree plot from principal component analysis

eFigure 2 shows a plot of the eigenvalues. The scree test and eigenvalue>1 citerion were originally developed for scenarios with few variables. In our case, there were millions of variables and even small eigenvalues corresponded to tens of thousands of sites. We therefore took a conservative approach and selected the first seven principal components.



eFigure 3. QQ plot

The QQ plot in e**Figure 3** shows that the distribution of the *p*-values from the MWAS. Because we used a log scale the majority of *p*-values will be in the lower range and fall on the straight line, indicating the expected *p*-value distribution (red line) under the null hypothesis that assumes no effects of the markers. However, there is also evidence that markers in the right upper corner of the plot have *p*-values smaller than would be expected under the null hypothesis, suggesting true association between these markers and the case/control status. These deviations are unlikely to occur by chance as they exceed the 95% confidence intervals (blue lines) for the null distribution. The plots also list the lambda values (i.e., the ratio of the median observed *p*-value of the distribution to the expected *p*-value under the null hypothesis).

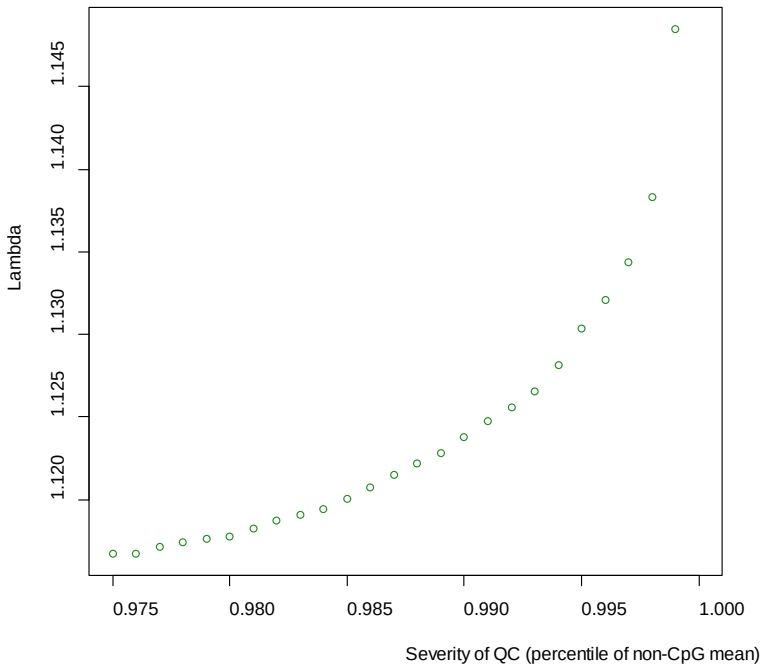


A table giving the specific location, test statistics, *p*-values and *q*-values for all MWAS findings that had $q < 0.01$ is provided in a separate excel spreadhseet.

eFigure 4. Relation between severity of block QC and lambda

eFigure 4 shows the relationship between severity of block QC and lambda. The blocks were QC'ed by eliminating low coverage blocks using increasingly higher thresholds. Blocks with higher coverage will 1) generally be more reliably measured (more reads translates to more precise methylation estimates) and 2) less likely to be unmethylated.

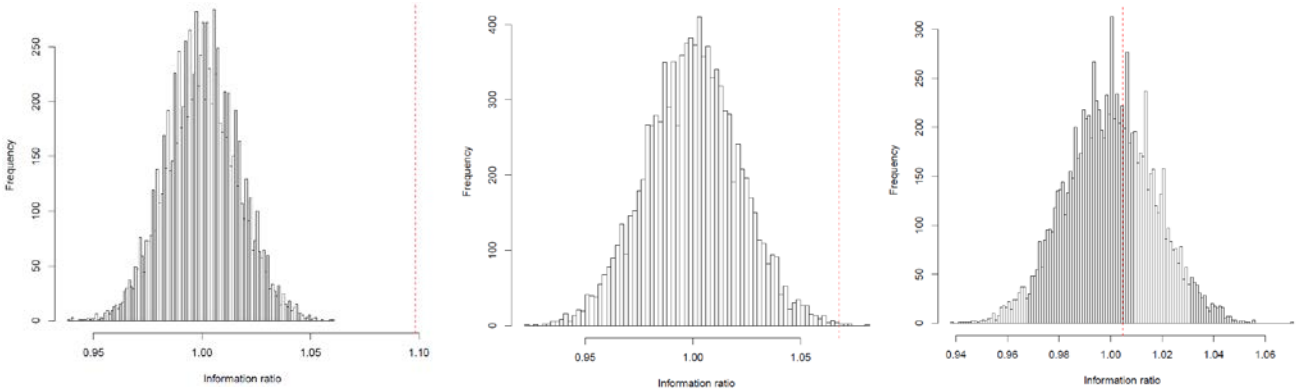
To determine the QC cut-off we first selected (non-CpG) sites that were at least 400bp away from the nearest CpG. As MBD-seq can only enrich for methylation occurring at CpGs, the coverage observed at these non-CpG sites cannot be due to methylation but reflects a "noise" coverage level. The x-axis shows the percentile of non-CpG coverage used to eliminate all



blocks that had a lower coverage.

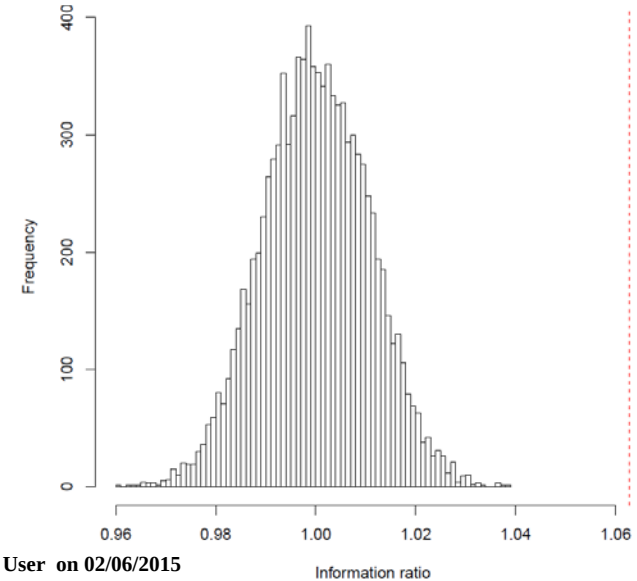
eFigure 5. Permutation test results for remaining hypoxia genes

a) Protein-protein interaction b) Pathway c) miRNA



To perform the permutation tests, we first removed the top MWAS findings that were used in the initial network analyses and then performed 10,000 permutations to test whether MWAS results for the remaining network genes had better *p*-values in the MWAS than expected under the null hypothesis. Our test statistic was an "information ratio" calculated as the observed number of methylation sites in hypoxia genes with a *p*-value < 0.05 divided by the number of sites with a *p*-value < 0.05 expected by chance.

The figures show a histogram where the frequency (y-axis) of the 10,000 information ratio test statistics obtained after random permutations is plotted against value of the information ratio test statistics (x-axis). The red vertical line marks the observed value of the information ratio. For both networks created using protein-protein interactions, none of the test statistics values obtained after permutation had a value more extreme then the observed test statistics. As we used 10,000 permutations, a *p*-value < 0.0001 (=1/10000) indicates that MWAS results for remaining group of network genes were more significant than expected under the null hypothesis. For the pathway analyses, the permutation test was also highly significant *p*-value < 0.0009. When we performed this test for microRNA mir-217 that, among other functions, regulates heme oxygenase 1, an enzyme responsive to hypoxic conditions[3], results were not significant .In this case the other genes regulated by mir-217 do not show evidence of being relevant for hypoxia (p-value < 0.39). This is a reasonable result as we would not expect mir-217 to only regulate hypoxia related genes.



eFigure 6. Permutation test results for remaining immune system genes

For the immune system, we combined protein-protein interactions and pathway results to avoid small sets of genes. We again observed that none of the test statistics obtained after permutation had a value more extreme than the observed test statistics. As we used 10,000 permutations this means the *p*-value is less than 1/10000, indicating that MWAS results for the remaining group of network genes were more significant than expected under the null hypothesis.

2. Replication

The replication was done in an independent sample from the same study population as the MWAS samples. DNA was extracted following the same procedure as was used for the MWAS samples. Genomic DNA was bisulfite converted using Epitect 96 (QIAGEN, Germantown, MD). For each assay, the bisulfite converted DNA was used as input material for the PyroMark PCR in a total reaction volume of 25 µl. Five µl of the amplicons were used in the PyroMark pyrosequencing reactions (QIAGEN, Germantown, MD). All procedures were conducted according to the standard protocols provided by the vendor.

Each pyrosequencing assay was designed using the PyroMark Assay Design software (QIAGEN, Germantown, MD). High quality assays targeting the selected regions were evaluated *in silico* by scanning for possible secondary binding sites throughout the bisulfite converted human genome using the BiSearch Web Server[4-5]. Furthermore, each assay was evaluated in the laboratory prior to investigating the full-scale study sample. The laboratory technical evaluation included four different negative controls, five standards with known methylation levels as well as eight samples from the independent sample. The evaluation was conducted using the same protocol as was used for the investigation of the complete replication sample with the exception of the controls where specific reagents were excluded. The negative pyrosequencing controls consisted of 1) PCR product conducted without DNA, replaced with H_2O, 2) pyrosequencing reaction without amplicon or PCR primers, 3) pyrosequencing reaction without amplicon but with the biotinlylated PCR primer included, 4) pyrosequencing reaction without the sequencing primer and without amplicon but with the biotinlylated PCR primer included. The known standards included five DNA samples in duplicates with known methylation levels (0%, 25%, 50%, 75% and 100% methylation) that were created using methylated (#59665) and unmethylated (#59655) EpiTect Control DNA (QIAGEN, Valencia, CA).

The evaluation of the assays included a rigorous quality control. First, all negative controls were checked to be truly negative. Next, the known standards were checked to have values in the approximate range of the expected methylation level. We occasionally noticed that, although values were linear, the methylation levels for the known standards were shifted for different CpG sites within the same assay or consistently slightly shifted for an entire assay. These assays were still evaluated as good quality assays, as the shifts are likely to be true observations, caused by slightly variable methylation levels in the synthetically created fully methylated and unmethylated DNA. Assays with large discrepancies to the expected values were rerun or/and redesigned. Finally, the pyrograms of the known standards and the eight individuals from the study sample were checked to match the reference sequence pattern and the peak heights were checked to match the histogram of expected values.

Assays that passed the evaluation were run for the replication sample. e**Table 2** shows all replication assays. Each 96-well sample plate contained four negative controls (PCR product conducted without DNA, replaced with H_2O) and two DNA samples with known methylation levels in duplicate (0% and 100% methylation). For each plate, the negative controls were checked to be negative and the known standards were checked to agree with the measure from

their previous methylation levels on the evaluation plate. Furthermore, the pyrogram for each individual was checked to match the reference sequence pattern and the expected height of the histograms. If discrepancies were observed the trailing sequence was excluded from further analysis. If failing or uncertain dispensation occurred, all trailing sequence was excluded. Finally, to ensure that the bisulfite conversion was sufficient we included bisulfite conversion controls in each assay.

eTable 2. Design features of pyrosequencing assays

		Location			Primers		
	Chr	Begin	End	Strand	Forward PCR	Reverse PCR	Sequencing
ARHGAP26	5	142138110	142138140	+	GAAGAAAGATTGTTAGGTTTGAAGTAAG	Bio-CATAACAATTAAATAACCCCAAACCA	AGGTTTGAAGTAAGAAATAATT
ARNT	1	150825039	150825079	+	AGAATTGATATAGAGGGTTTGTAAAT	Bio-ACTTCATCCTAAAACTATAAATATTCATAT	TAGAGGGTTTGTAAATG
ATXN1	6	16295927	16295927	+	ATAAGTTATTTTAGGAGGGGTAAGGA	Bio-CCAATTAATTACTTTTATTCTTTTACAACA	GTAAGGAGAGGAGGAT
CREB1	2	208461648	208461657	+	AAGGTAATTTAAAATAAATATGTGGAAAT	Bio-ACTTTTAACTCCTCAATCAATATCTTAT	GTTTGTAGGGAAGTAGTT
CTAGE11P	13	75795508	75795510	+	GTATGGGTTTGGGTTTTTAGTTAT	Bio-AACCACTATTCCTAATCCCTACTTTATTCT	GGTTTGGGTTTTTAGTTATT
FCAR	19	55383906	55384107	-	Bio-AAATGAAATAGTTAATATGGAAGTTGAT	CCCACTTTATTTTCATTTTAATACATCA	CATAACTTAACCCTTCCAA
ETS2	21	40183132	40183132	+	Bio-TTGATGTGTTGAAATAGGAAGTATAGAT	TTCCCACAAACATTACCCAAATAACTAAA	CCCAAATAACTAAAAAATACCTT
FAM63B	15	59146738	59146756	+	GTGGAAGATAATTTGGGAATAGTGA	Bio-TCCAACAAAAACCAACACTTATTACA	AAGATAATTTGGGAATAGTGAA
FNDC3B	3	171787965	171788032	+	AGGGTATGTATGATGTTTTAAGGT	Bio-CAAATCTCAATTTAACACCAAAACATCTA	TTTAATATGTTAATATTTTTTTAGG
HTRA3	4	8281202	8281234	+	GGAGGTAGTAGTGTTGTGTAGATT	Bio-TCCACCCCCTACTTACCAA	GGTAGTAGTGTTGTGTAGATTT
Inter-genic	16	54209779	54209801	-	Bio-AAATGTTAATAGTTTTAGTAAGGAATATTG	CCCCTCTTTCTTTCTTTTCTTTAAATAT	AAATAAAAAAATTTCTTCTTCCA
PARK2	6	162054839	162054989	-	AAGTTGTTTGGGTTTTAGGG	Bio-AAAAAACCTCTATCCAAAATATCTCC	ATTTGGTTAAAGTTTTTTTTTTTATG
PPARA	22	46655187	46655187	+	TAGGATAATTTTTGGAGTTTATTGAATGT	Bio-TACCAAAAATAACTTTCTATTAATACCACT	GTTAGAAGTTTATTTGGATGGG
RCOR1	14	103130134	103130147	+	AGGGTAGTGTTAAGAGAAATAGG	Bio-ACCAAAAACACTTTTCTTCCATATTATC	AAGTGTTTTGTTTTTGTTTAG
RELN	7	103576641	103576649	-	AGTTTAAGATTAGTTTGGTTAATATGTTGA	Bio-CCCTATTCATCTTACATAAATTACTTCA	AATTAGTTGGGTATGGT
SATB1	3	18443092	18443135	+	AGTGGTTGGTTTGTATAAAATATTAATGT	Bio-TTATAAAACCCACAACCAATTTCTACC	AATGTTTATAGAATAAAATAAATT
SMAD3	15	67396559	67396564	-	Bio-TGTGATGAGAATTGGGAAGGTA	CCCAACTTTTAACAACCAAAAACTTTTC	AATATATTTCAACTTCTATAATC
TBC1D22A	22	47170345	47170371	+	GGGGAATTTTTATTTTGAGTATTTTTAGAT	Bio-TCTCAACCTCTACTTCCTACTTAATA	TTTGGTGATAAATATGGAAATT

Note: The locations are given for the CpGs targeted in each assay. Strand indicates if the assays were designed on the upper (+) or the lower (-) strand. The three primers used for each assay are given. The direction of the sequencing primer is always the opposite of the biotinylated primer.

eTable 3 gives the full replication results. In order to test if there was an association between case\control status and methylation level, we performed multiple logistic regression with age, sex and plate indicator variables as covariates. Additionally, outlying observations that were more than five standard deviations away from the mean level of methylation for a specific CpG site were removed from the analysis so that the results would not be unduly swayed by these observations.

eTable 3. Full replication results

Gene	Chr	Position (bp)	n	Beta	T-value	P-value
ARHGAP26	5	142138110	333	-0.044	-2.97	2.96E-03
		142138121	331	-0.041	-2.56	1.03E-02
		142138140	315	-0.047	-2.77	5.54E-03
ARNT	1	150825039	1087	-0.048	-6.05	1.43E-09
		150825079	1011	-0.043	-3.36	7.87E-04
ATXN1	6	16295927	1120	0.001	0.05	9.61E-01
		16295949	1089	0.015	0.88	3.81E-01
CREB1	2	208561648	1100	-0.047	-6.34	2.33E-10
		208561657	1086	-0.048	-6.46	1.03E-10
CTAGE11P	3	75795508	344	-0.045	-3.00	2.71E-03
		75795510	330	-0.039	-2.65	8.04E-03
ETS2	21	40183132	349	-0.047	-2.88	3.98E-03
FAM63B	15	59146738	1075	-0.054	-6.95	3.76E-12
		59146744	1065	-0.052	-6.84	7.72E-12
		59146756	1033	-0.051	-6.34	2.31E-10
FCAR	19	55384107	1093	-0.037	-2.72	6.61E-03
		55384104	1058	-0.033	-2.20	2.75E-02
FNDC3B	3	171788002	189	-0.056	-3.30	9.58E-04
		171788032	188	-0.057	-2.91	3.65E-03
HTRA3	4	8281202	344	0.001	0.12	9.07E-01
		8281204	339	0.002	0.20	8.40E-01
Inter-genic	16	54209779	351	-0.025	-1.75	8.07E-02
		54209801	353	-0.025	-1.97	4.85E-02
PARK2	6	162054935	333	0.043	0.64	5.24E-01
		162054989	330	0.019	1.00	3.18E-01
PPARA	22	46655187	1094	-0.047	-6.19	5.94E-10
		232035572	305	-0.038	-2.48	1.31E-02
RCOR1	14	103130134	1082	-0.049	-6.43	1.26E-10
		103130138	1069	-0.046	-5.58	2.35E-08
		103130147	1027	-0.011	-0.62	5.34E-01
RELN	7	103576688	1017	0.068	2.95	3.14E-03
		103576649	944	0.041	3.42	6.28E-04
SATB1	3	18443092	330	-0.034	-2.16	3.08E-02
		18443106	329	-0.032	-1.91	5.64E-02
		18443135	327	-0.036	-2.08	3.71E-02
SMAD3	15	67396559	1109	-0.086	-7.66	1.81E-14
		67396564	1099	-0.079	-7.49	6.82E-14
TBC1D22A	22	47170345	335	-0.050	-2.73	6.30E-03
		47170354	328	-0.049	-2.32	2.03E-02

		47170371	323	-0.051	-2.14	3.24E-02

As the topic presents specific challenges and will be addressed in a separate paper, sex was regressed out from the MWAS. Indeed, when we tested for sex differences in our replication data, except for FNDC3B that reached nominal significant levels of p<0.004, none of the sex effects were significant.

Regressing out covariates

eTable 4 shows the correlations between smoking status (subjects yes/no smoked) and the methylation of hypoxia genes plus results of regression analyses where smoking was and was not included as a covariate.

eTable 4. Smoking and hypoxia: Bivariate correlations and regression analyses

			Cor.	Smoking *not* regressed out			Smoking regressed out		
Gene	Chr	Position (bp)		N	Beta	P-value	N	Beta	P-value
ARHGAP26	5	142138110	-0.02	333	-0.044	2.96E-03	306	-0.058	6.77E-04
		142138121	-0.016	331	-0.041	1.03E-02	305	-0.051	6.21E-03
		142138140	-0.036	315	-0.047	5.54E-03	288	-0.060	1.91E-03
ARNT	1	150825039	-0.022	1087	-0.048	1.43E-09	883	-0.062	9.12E-10
		150825079	-0.026	1011	-0.043	7.87E-04	821	-0.039	1.42E-02
CREB1	2	208561648	-0.069	1100	-0.047	2.33E-10	895	-0.046	3.94E-07
		208561657	-0.062	1086	-0.048	1.03E-10	888	-0.048	1.17E-07
ETS2	21	40183132	-0.032	349	-0.047	3.98E-03	320	-0.055	2.82E-03
SATB1	3	18443092	-0.011	330	-0.034	3.08E-02	302	-0.039	3.11E-02
		18443106	-0.025	329	-0.032	5.64E-02	300	-0.040	3.44E-02
		18443135	-0.039	327	-0.036	3.71E-02	298	-0.044	2.51E-02
SMAD3	15	67396559	-0.116*	1109	-0.086	1.81E-14	904	-0.091	6.90E-11
		67396564	-0.093*	1099	-0.079	6.82E-14	894	-0.080	1.26E-09

Cor is Pearson correlation, N is sample size, Beta is standardized regression coefficient for case-control status. * is P < 0.01.

Results show that smoking is uncorrelated with the methylation of hypoxia genes. A possible exception is SMAD3 that shows a correlation of 0.1. The regression analyses show that smoking does not account for the associations between case-control status and the methylation of hypoxia genes. As there were missing observations for the smoking variable, there are slightly higher P values. However, the better beta-coefficients suggest that to the extent these changes are meaningful, regressing out the effects of smoking increases the replication evidence for the hypoxia genes.

We also conducted a MWAS in controls only with smoking status as the outcome variable. Even when we used a very liberal *q*-value threshold of 0.75, none of the hypoxia genes were found in the top results.

eTable 5 and **6** show similar analyses for alcohol and use of narcotics. The regression analyses again show that these covariates do not account for the associations between case-control status and the methylation of hypoxia genes.

eTable 5. Alcohol: Bivariate correlations and regression analyses

			Cor.	Alcohol *not* regressed out			Alcohol regressed out		
Gene	Chr	Position (bp)		N	Beta	P-value	N	Beta	P-value
ARHGAP26	5	142138110	-0.017	333	-0.044	2.96E-03	298	-0.045	1.06E-02
		142138121	-0.037	331	-0.041	1.03E-02	296	-0.039	4.04E-02
		142138140	-0.024	315	-0.047	5.54E-03	284	-0.051	1.10E-02
ARNT	1	150825039	-0.036	1087	-0.048	1.43E-09	1019	-0.048	1.54E-07
		150825079	-0.044	1011	-0.043	7.87E-04	944	-0.046	2.69E-03
CREB1	2	208561648	-0.058	1100	-0.047	2.33E-10	1033	-0.045	1.19E-07
		208561657	-0.075	1086	-0.048	1.03E-10	1019	-0.043	4.53E-07
ETS2	21	40183132	-0.065	349	-0.047	3.98E-03	313	-0.046	2.04E-02
SATB1	3	18443092	-0.044	330	-0.034	3.08E-02	298	-0.036	5.36E-02
		18443106	-0.021	329	-0.032	5.64E-02	297	-0.037	6.59E-02
		18443135	-0.038	327	-0.036	3.71E-02	295	-0.033	1.14E-01
SMAD3	15	67396559	-0.015	1109	-0.086	1.81E-14	1039	-0.085	7.41E-11
		67396564	-0.028	1099	-0.079	6.82E-14	1030	-0.081	4.01E-11

eTable 6. Narcotics: Bivariate correlations and regression analyses

			Cor.	Narcotics *not* regressed out			Narcotics regressed out		
Gene	Chr	Position (bp)		N	Beta	P-value	N	Beta	P-value
ARHGAP26	5	142138110	-0.020	333	-0.044	2.96E-03	332	-0.043	3.49E-03
		142138121	-0.016	331	-0.041	1.03E-02	330	-0.040	1.40E-02
		142138140	-0.036	315	-0.047	5.54E-03	314	-0.046	6.45E-03

ARNT	1	150825039	0.002	1087	-0.048	1.43E-09	1081	-0.047	5.69E-09
		150825079	-0.003	1011	-0.043	7.87E-04	1005	-0.043	1.21E-03
CREB1	2	208561648	-0.038	1100	-0.047	2.33E-10	1094	-0.045	1.91E-09
		208561657	-0.036	1086	-0.048	1.03E-10	1080	-0.046	9.84E-10
ETS2	21	40183132	-0.032	349	-0.047	3.98E-03	348	-0.044	6.97E-03
SATB1	3	18443092	-0.011	330	-0.034	3.08E-02	329	-0.034	3.31E-02
		18443106	-0.025	329	-0.032	5.64E-02	328	-0.032	5.61E-02
		18443135	-0.039	327	-0.036	3.71E-02	326	-0.036	3.84E-02
SMAD3	15	67396559	-0.114*	1109	-0.086	1.81E-14	1103	-0.085	5.69E-14
		67396564	-0.078	1099	-0.079	6.82E-14	1094	-0.080	1.24E-13

References

1. Bock, C., Tomazou, E.M., Brinkman, A.B., Muller, F., Simmer, F., Gu, H., Jager, N., Gnirke, A., Stunnenberg, H.G., and Meissner, A. (2010). Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol 28, 1106-1114.
2. Chavez, L., Jozefczuk, J., Grimm, C., Dietrich, J., Timmermann, B., Lehrach, H., Herwig, R., and Adjaye, J. (2010). Computational analysis of genome-wide DNA methylation during the differentiation of human embryonic stem cells along the endodermal lineage. Genome Res 20, 1441-1450.
3. Beckman, J.D., Chen, C., Nguyen, J., Thayanithy, V., Subramanian, S., Steer, C.J., and Vercellotti, G.M. (2011). Regulation of heme oxygenase-1 protein expression by miR-377 in combination with miR-217. J Biol Chem 286, 3194-3202.
4. Tusnady, G.E., Simon, I., Varadi, A., and Aranyi, T. (2005). BiSearch: primer-design and search tool for PCR on bisulfite-treated genomes. Nucleic Acids Res 33, e9.
5. Aranyi, T., Varadi, A., Simon, I., and Tusnady, G.E. (2006). The BiSearch web server. BMC Bioinformatics 7, 431.