The American Journal of Human Genetics, Volume *98*

Supplemental Data

DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis

Bonnie R. Joubert, Janine F. Felix, Paul Yousefi, Kelly M. Bakulski, Allan C. Just, Carrie Breton, Sarah E. Reese, Christina A. Markunas, Rebecca C. Richmond, Cheng-Jian Xu, Leanne K. Küpers, Sam S. Oh, Cathrine Hoyo, Olena Gruzieva, Cilla Söderhäll, Lucas A. Salas, Nour Baïz, Hongmei Zhang, Johanna Lepeule, Carlos Ruiz, Symen Ligthart, Tianyuan Wang, Jack A. Taylor, Liesbeth Duijts, Gemma C. Sharp, Soesma A. Jankipersadsing, Roy M. Nilsen, Ahmad Vaez, M. Daniele Fallin, Donglei Hu, Augusto A. Litonjua, Bernard F. Fuemmeler, Karen Huen, Juha Kere, Inger Kull, Monica Cheng Munthe-Kaas, Ulrike Gehring, Mariona Bustamante, Marie José Saurel-Coubizolles, Bilal M. Quraishi, Jie Ren, Jörg Tost, Juan R. Gonzalez, Marjolein J. Peters, Siri E. Håberg, Zongli Xu, Joyce B. van Meurs, Tom R. Gaunt, Marjan Kerkhof, Eva Corpeleijn, Andrew P. Feinberg, Celeste Eng, Andrea A. Baccarelli, Sara E. Benjamin Neelon, Asa Bradman, Simon Kebede Merid, Anna Bergström, Zdenko Herceg, Hector Hernandez-Vargas, Bert Brunekreef, Mariona Pinart, Barbara Heude, Susan Ewart, Jin Yao, Nathanaël Lemonnier, Oscar H. Franco, Michael C. Wu, Albert Hofman, Wendy McArdle, Pieter Van der Vlies, Fahimeh Falahi, Matthew W. Gillman, Lisa F. Barcellos, Ashish Kumar, Magnus Wickman, Stefano Guerra, Marie-Aline Charles, John Holloway, Charles Auffray, Henning W. Tiemeier, George Davey Smith, Dirkje Postma, Marie-France Hivert, Brenda Eskenazi, Martine Vrijheid, Hasan Arshad, Josep M. Antó, Abbas Dehghan, Wilfried Karmaus, Isabella Annesi-Maesano, Jordi Sunyer, Akram Ghantous, Göran Pershagen, Nina Holland, Susan K. Murphy, Dawn L. DeMeo, Esteban G. Burchard, Christine Ladd-Acosta, Harold Snieder, Wenche Nystad, Gerard H. Koppelman, Caroline L. Relton, Vincent W.V. Jaddoe, Allen Wilcox, Erik Melén, and Stephanie J. London

Figure S1. Comparison of –log10(p values) from the model evaluating the effect of sustained maternal smoking during pregnancy on methylation (primary model) to the same model additionally adjusted for cell type proportion indicated strongly correlated p values (correlation coefficient = 0.92 across all CpGs, 0.98 for the CpGs FDR significant in the primary model).

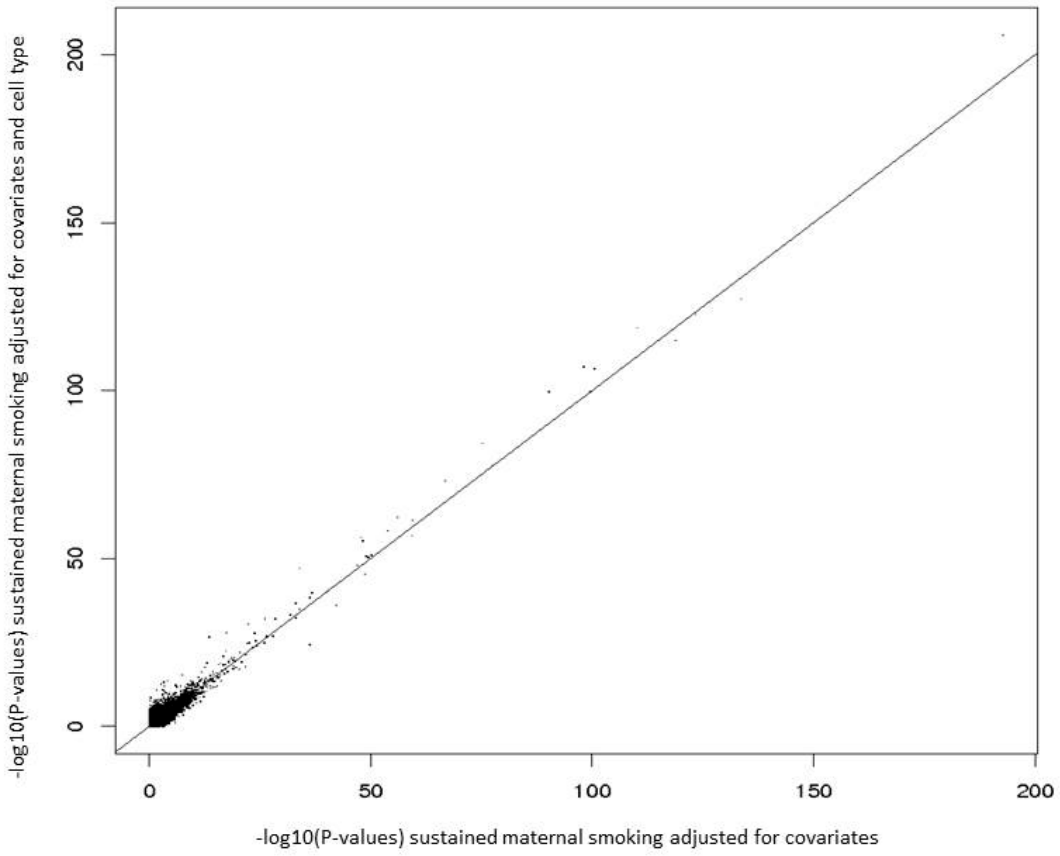

Figure S2. Top 15 statistically enriched gene ontology, biological processes, among genes differentially methylated with sustained smoke exposure. Enrichment was tested using Fisher's exact test and the topGO R package.

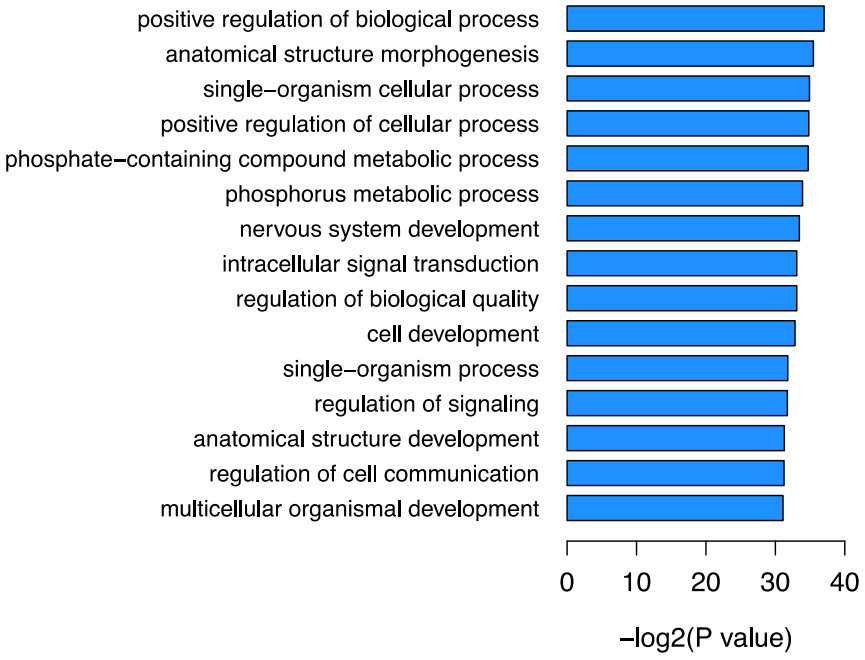

Figure S3. Enriched gene ontology biological processes reaching Benjamini-Hochberg threshold of 0.1, among genes differentially methylated with sustained smoke exposure. Enrichment was tested using the DAVID bioinformatics resource.

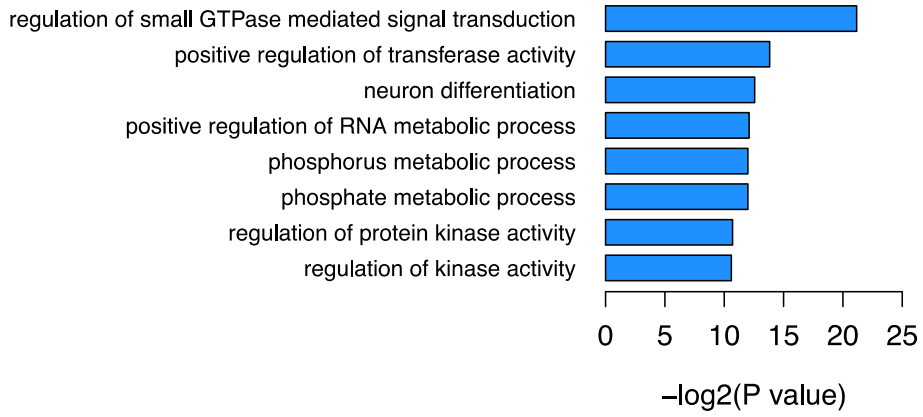

Figure S4**.** Top 15 statistically enriched diseases and biofunctions in genes differentially methylated with sustained smoke exposure. Enrichment was tested using Qiagen's IPA software.

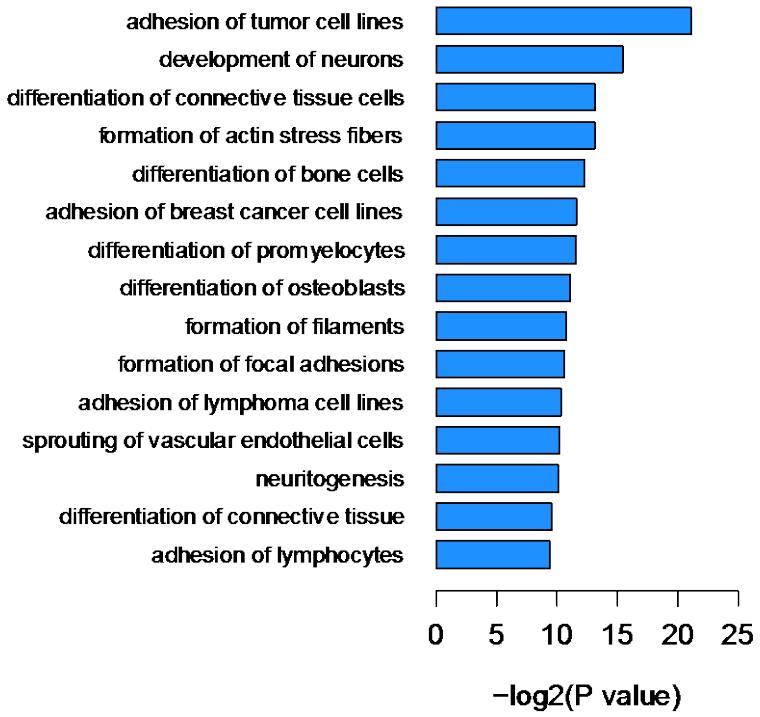

Figure S5. Meta-analysis of the association between any maternal smoking and DNA methylation in newborn cord blood. A total of 4,653 CpGs were considered statistically significant using FDR correction (solid horizontal line); 407 Bonferroni significant (dashed horizontal line).

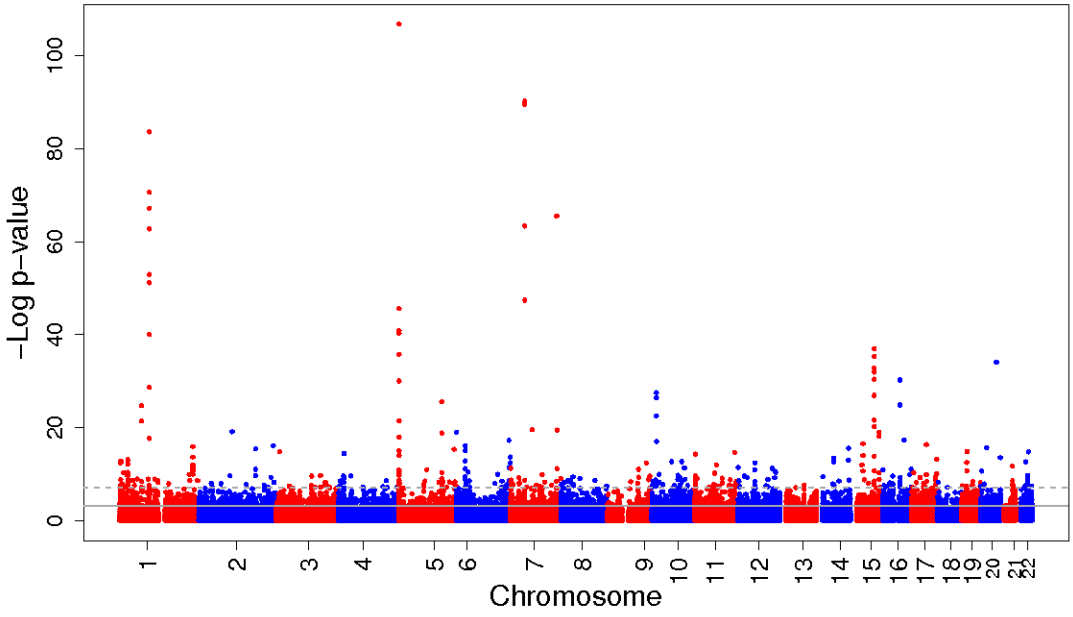

Figure S6. Comparison of –log10(p values) from the model evaluating effect of sustained maternal smoking during pregnancy on methylation (primary model) with the model evaluating the effect of any maternal smoking during pregnancy on methylation (correlation coefficient = 0.68 across all CpGs, 0.96 across CpGs FDR significant in the primary model).

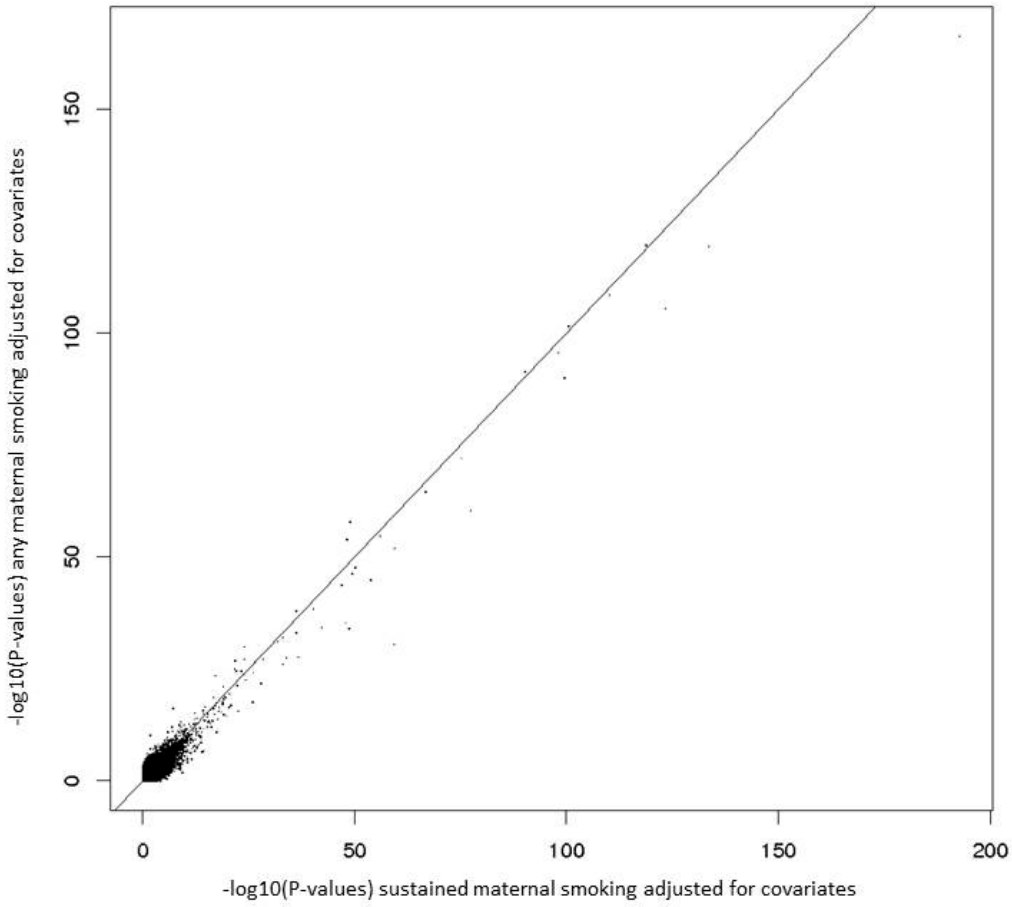

Figure S7. Meta-analysis results for the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood: CpGs in or near *BMP4*.
Panel A: -log10(P values) from the meta-analysis model, CpGs indicated by dots, color coded based on pairwise correlation with neighboring CpGs. Panel B: Annotation tracks for the plotted genomic region. Panel C: Pairwise correlation matrix across the displayed CpGs.

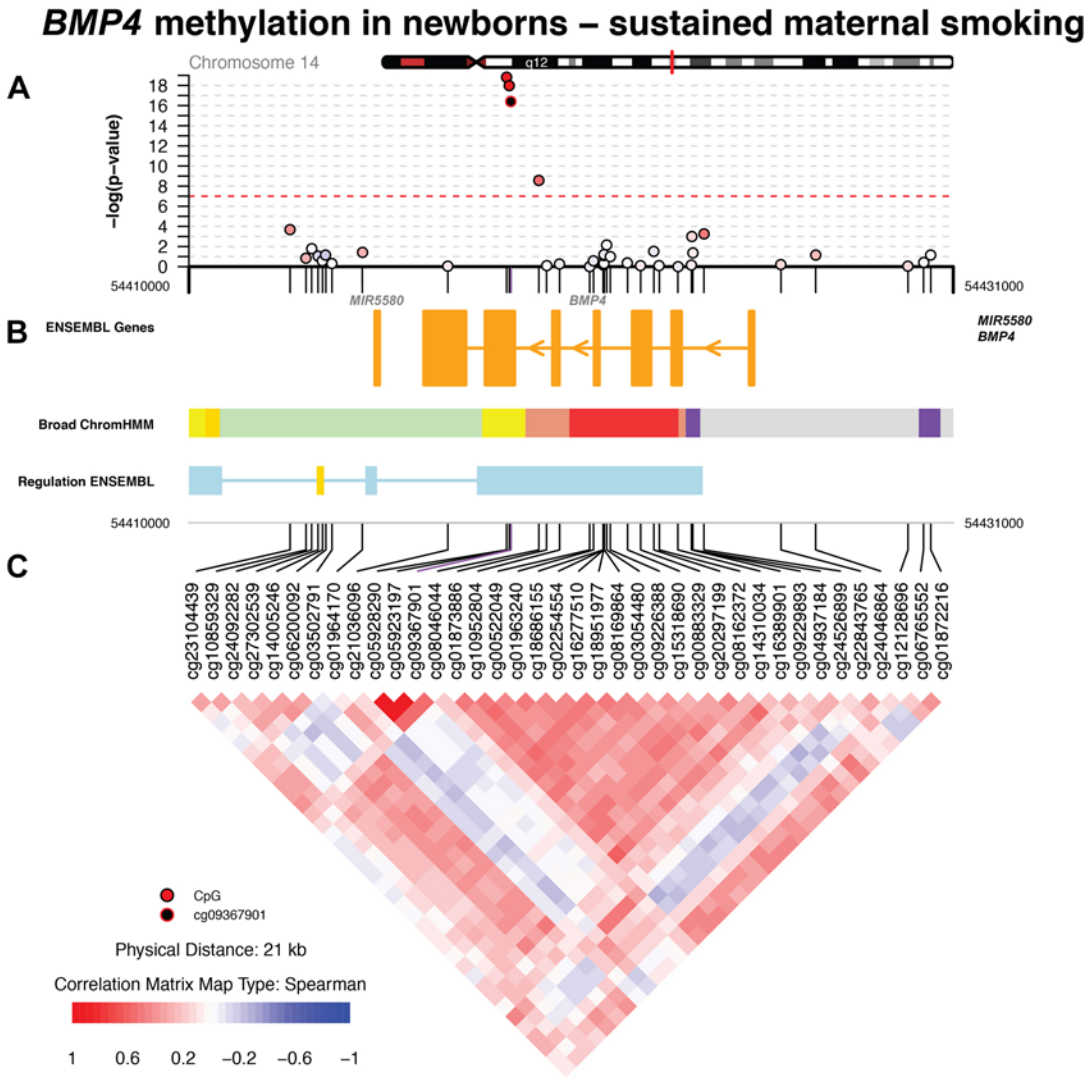

Figure S8. Meta-analysis results for the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood: CpGs in or near *BHMT2*.
Panel A: -log10(p values) from the meta-analysis model, CpGs indicated by dots, color coded based on pairwise correlation with neighboring CpGs. Panel B: Annotation tracks for the plotted genomic region. Panel C: Pairwise correlation matrix across the displayed CpGs.

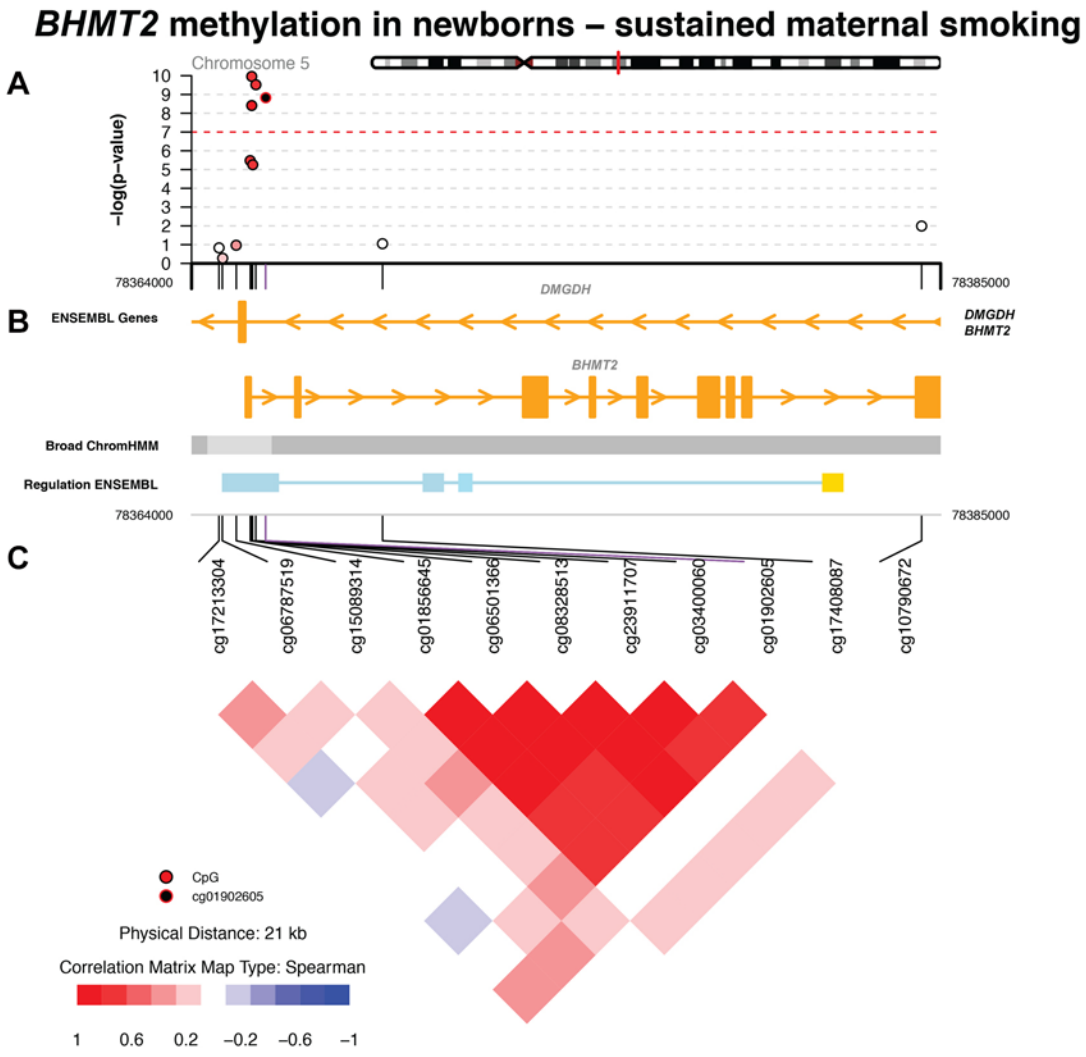

Figure S9. Meta-analysis results for the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood: CpGs in or near *IL32*.
Panel A: -log10(p values) from the meta-analysis model, CpGs indicated by dots, color coded based on pairwise correlation with neighboring CpGs. Panel B: Annotation tracks for the plotted genomic region. Panel C: Pairwise correlation matrix across the displayed CpGs.

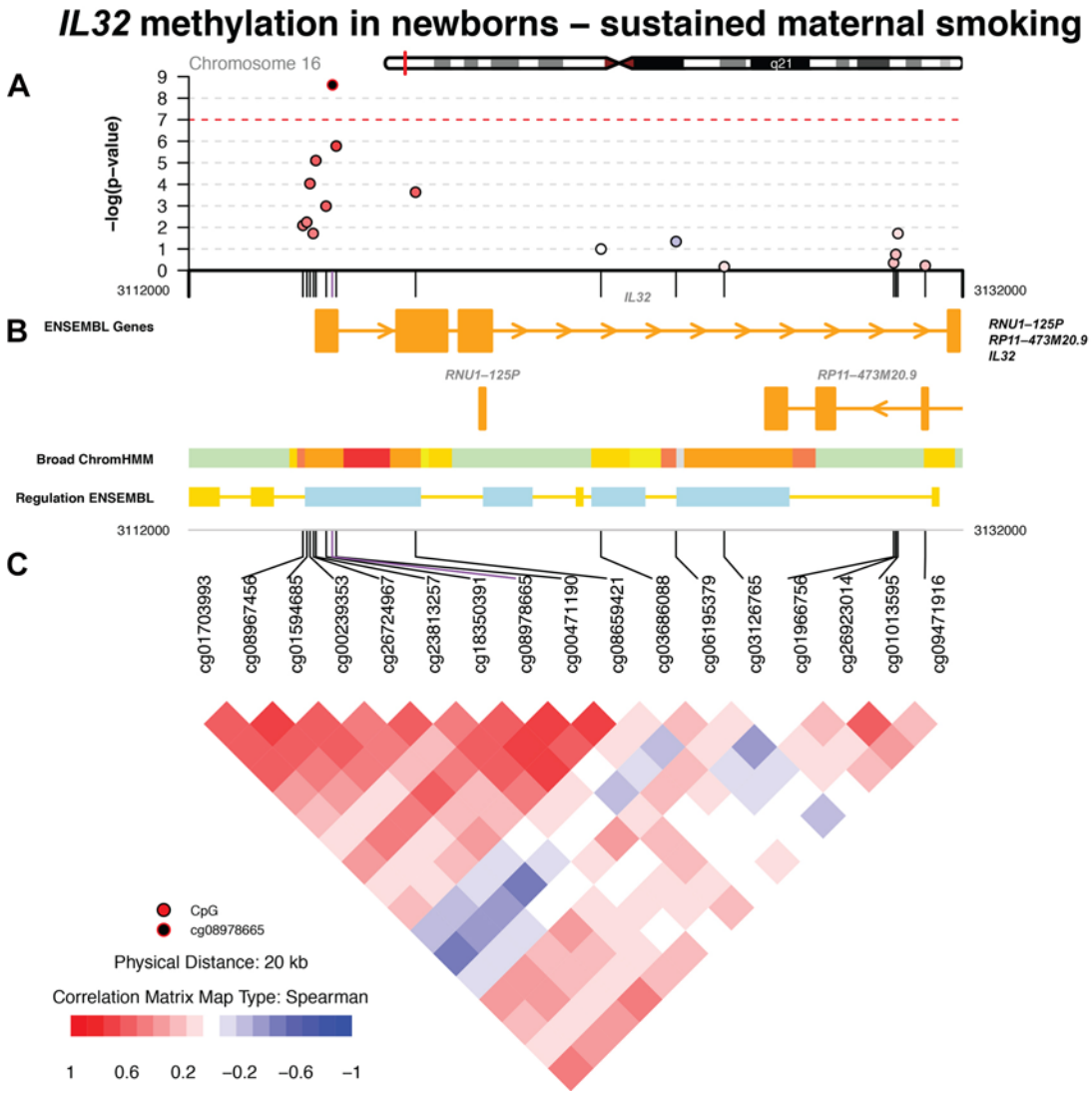

Table S1. Cohort-specific covariate frequencies

Cohort	Study population	Ancestry (%)	Maternal age (years) Mean (SD)	Maternal education N (%)				Parity N (%)			
				Level 1	Level 2	Level 3	Level 4	0	1	2	3 or more
ALSPAC	newborns	European	30.0 (4.4)	141 (16.4)	293 (34.1)	248 (28.8)	178 (20.7)	408 (47.4)	311 (36.2)	109 (12.7%)	32 (3.7)
CHAMACOS	newborns	Mexican-American	25.2 (5.0)	301 (79.6)	44 (11.6)	31 (8.2)	2 (0.5)	135 (35.7)	109 (28.8)	80 (21.2)	54 (14.3)
CHS	newborns	European	29.3 (5.8)	9 (11.4)	11 (13.9)	36 (32.9)	23 (21.9)	25 (29.4)	31 (36.5)	23 (27.1)	6 (7.1)
GECKO	newborns	European	30.3 (0.26)	175 (68.6)	80 (31.4)	-	-	87 (34.1)	116 (45.5)	40 (15.7)	12 (4.7)
Generation R	newborns	European	31.5 (4.2)	19 (2.2)	301 (34.1)	226 (25.7)	333 (37.9)	525 (59.5)	274 (31.0)	72 (8.2)	12 (1.4)
IOW	newborns	European	21.42 (1.31)	2 (2.2)	32 (35.5)	24 (26.6)	32 (35.5)	50 (55.5)	28 (31.1)	7 (8.0)	2 (2.0)
MeDALL	newborns	European	30.7 (4.6)	59 (16.3)	183 (50.6)	108 (29.8)	-	177 (48.9)	161 (44.5)	12 (3.3)	1 (2.7)
MOBA1	newborns	European	29.9 (4.3)	78 (7.3)	344 (32.4)	471 (44.3)	170 (16.0)	434 (40.8)	435 (40.9)	138 (13.0)	56 (5.3)
MOBA2	newborns	European	30.0 (4.5)	54 (8.0)	224 (33.4)	269 (40.1)	124 (18.5)	275 (41.0)	230 (34.3)	137 (20.4)	29 (4.3)
MOBA3	newborns	European	29.8 (4.4)	21 (10.1)	43 (20.8)	104 (50.2)	39 (18.8)	105 (50.7)	68 (32.9)	25 (12.1)	9 (4.3)
NFCS	newborns	European	29.6 (4.9)	118 (13.3)	421 (47.4)	299 (33.6)	51 (5.7)	367 (41.3)	311 (35.0)	156 (17.6)	55 (6.2)
NEST	newborns	European (49) African American (45) Other (6)	28.9 (6.4)	47 (11.2)	92 (21.9)	115 (27.4)	166 (39.5)	136 (32.9)	135 (32.6)	83 (20.0)	60 (14.5)
Project Viva	newborns	European (71) African American (12) Hispanic (8) Asian (5) Other (5)	32.1 (5.3)	11 (2.3)	33 (6.8)	119 (24.5)	322 (66.4)	225 (46.4)	177 (36.5)	59 (12.2)	24 (4.9)
ALSPAC	older children	European	30.2 (4.3)	123 (14.6)	291 (34.6)	249 (29.6)	177 (21.1)	389 (46.3)	313 (37.3)	106 (12.6)	32 (3.8)
BAMSE	older children	European	30.8 (4.3)	28 (8.5)	82 (24.8)	102 (30.7)	119 (36)	185 (55.9)	106 (32.0)	40 (12.1)	-
GALA II	older children	Mexican (49) Puerto Rican (38) Other Latino (14)	25.3 (6.3)	223 (39.2)	153 (26.9)	119 (20.9)	69 (12.1)	270 (47.5)	162 (28.5)	83 (14.6)	54 (9.5)
MeDALL	older children	European	30.8 (4.3)	197 (23.1)	340 (40.0)	301 (35.4)	-	430 (50.5)	347 (40.8)	52 (6.1)	11 (1.2)
SEED	older children	European (58) African American (12) Other (30)	36.6 (5.3)	23 (3.9)	60 (10.3)	155 (26.5)	346 (59.2)	-	-	-	-

[a] Frequencies for categories displayed for descriptive purposes. Some cohorts collapsed covariate categories in statistical models to accommodate small sample size. Maternal education categories were standardized to capture European and North American education systems (Level 1: less than secondary school; Level 2: secondary school completion; Level 3: some college or university; Level 4: college degree or more). A few studies had insufficient data for some covariates or specific categories, indicated with "-."

Table S2. Cohort and model-specific lambdas and probe number

Study	Study Population	Lambda Values				Number of probes	
		Sustained smoking, raw betas	Sustained smoking, normalized betas	Any smoking, raw betas	Any smoking, normalized betas	Raw betas	Normalized betas
ALSPAC	newborns	1.66	1.09	1.80	1.03	485,512	485,512
CHAMACOS	newborns	NA	NA	1.05	1.00	435,369	435,369
CHS	newborns	NA	NA	2.27	1.41	383,857	482,650
GECKO	newborns	0.93	1.02	1.08	1.24	465,891	465,891
Generation R	newborns	1.43	1.56	1.96	1.78	436,013	436,010
IOW	newborns	NA	NA	4.44	0.83	485,577	343,203
MeDALL	newborns	0.84	1.16	1.04	1.54	485,512	439,306
MOBA1	newborns	1.31	1.27	1.31	1.07	473,844	473,844
MOBA2	newborns	1.48	0.81	1.16	1.33	473,748	473,748
MOBA3	newborns	0.90	0.95	1.00	1.29	466,629	466,629
NFCS	newborns	1.16	1.16	0.98	1.13	483,859	485,577
NEST	newborns	1.06	1.06	1.15	1.15	469,119	469,119
Project Viva	newborns	NA	NA	2.31	1.35	464,952	464,952
Meta-Analysis	newborns	2.16	1.47	2.96	1.36	464,696	464,628
ALSPAC	older children	-	0.79	-	-	-	485,512
BAMSE	older children	-	0.96	-	-	-	438,713
GALA II	older children	-	0.89	-	-	-	321,503
MeDALL	older children	-	0.94	-	-	-	439,306
SEED	older children	-	1.07	-	-	-	485,512
Meta-Analysis	older children	-	0.90	-	-	-	464,696

[a] Sustained smoking refers to models evaluating sustained maternal smoking during pregnancy as the exposure; any smoking refers to models evaluating any smoking by the mother during pregnancy as the exposure. Raw betas represent methylation values after quality control that were not normalized or subjected to additional data processing. Normalized betas represent methylation values after quality control that were normalized and/or processed according to cohort-specific protocols described in the Supplemental Methods. NA values indicate the model was not run for that cohort due to having less than 15 exposed subjects.

Table S4. Studies identified in literature review reporting statistically significant associations between smoking (maternal/personal) exposure and DNA methylation, after correction for multiple testing (FDR or Bonferroni)

First author	Year	Exposure	PMID
Breitling	2011	Adult smoking	21457905
Philibert	2012	Adult smoking	23070629
Siedlinski	2012	Adult smoking	22617718
Wan	2012	Adult smoking	22492999
Philibert	2013	Adult smoking	24120260
Zeilinger	2013	Adult smoking	23691101
Sun	2013	Adult smoking	23657504
Shenker	2013	Adult smoking	23175441
Dogan	2014	Adult smoking	24559495
Elliot	2014	Adult smoking	24485148
Besingi	2014	Adult smoking	24334605
Philibert	2014	Adult smoking	24120261
Tsaprouni	2014	Adult smoking	25424692
Wan	2014	Adult smoking	25517428
Zaghlool	2015	Adult smoking	25663950
Flanagan	2015	Adult smoking	25371448
Guida	2015	Adult smoking	25556184
Joubert	2012	Maternal smoking	22851337
Breton	2014	Maternal smoking	24964093
Markunas	2014	Maternal smoking	24906187
Harlid	2014	Maternal smoking	24704585
Lee	2014	Maternal smoking	25325234
Chhabra	2014	Maternal smoking	25482056
Maccani	2014	Maternal smoking	24283877
Richmond	2014	Maternal smoking	25552657
Ivorra	2015	Maternal smoking	25623364

Table S7. Probes flagged as potentially polymorphic or cross-reactive* among the 6,073 genome wide significant CpGs differentially methylated in newborn DNA in relation to maternal smoking

Chromosome	Position	CpG	Mapped Gene	Nearest Gene (10 Mb)	Diptest p value
1	156161651	cg24849049	-	*SLC25A44*	0.995
2	9471179	cg06627617	*ASAP2*	*ASAP2*	0.990
3	158465384	cg16757990	-	*RARRES1*	0.986
3	159563158	cg12847013	*SCHIP1*	*IQCJ-SCHIP1*	0.997
5	79549315	cg16518115	*SERINC5*	*SERINC5*	0.995
6	30095517	cg20999347	-	*TRIM40*	0.995
6	31804172	cg11931646	*C6orf48;SNORD52*	*C6orf48*	0.964
10	44231015	cg19730699	-	*HNRNPA3P1*	0.001
10	105219172	cg00453258	*CALHM1*	*CALHM1*	0.098
11	910094	cg07066326	*CHID1*	*CHID1*	0.730
11	2846932	cg17416793	*KCNQ1*	*KCNQ1*	0.913
11	124791601	cg09176023	*HEPACAM*	*HEPACAM*	0.719
12	20576950	cg03618302	*PDE3A*	*PDE3A*	0.990
14	104758477	cg20279254	-	*KIF26A*	0.908
16	1584118	cg08296037	*IFT140;TMEM204*	*IFT140*	0.991
17	14106800	cg13619177	*COX10*	*COX10*	0.984
17	74069256	cg21885995	*SRP68*	*SRP68*	0.991
18	32824104	cg23785882	*ZNF397*	*ZNF397*	0.979
19	58086380	cg23012294	*ZNF416*	*ZNF416*	0.759

[a] As identified by Chen et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203-9 (2013). Column headers: Mapped gene represents UCSC RefGene from Illumina annotation; Nearest Gene captures genes within 10 Mb using snipper software for annotation; Diptest p value computed using the diptest to evaluate variation from unimodality. Visual inspection of CpGs resulted in the above list of flags based on potential outlier values.

Supplemental Note

Cohort-specific methods; cohorts listed in alphabetical order

ALSPAC

ALSPAC design and study population

ALSPAC is a large, prospective cohort study based in the South West of England. 14,541 pregnant women resident in Avon, UK with expected dates of delivery 1st April 1991 to 31st December 1992 were recruited and detailed information has been collected on these women and their offspring at regular intervals.[1,2] The study website contains details of all the data that is available through a fully searchable data dictionary (http://www.bris.ac.uk/alspac/researchers/data-access/data-dictionary/).

As part of the ARIES (Accessible Resource for Integrated Epigenomic Studies, http://www.ariesepigenomics.org.uk/) project, the Infinium HM450 BeadChip has been used to generate epigenetic data on 1,018 mother-offspring pairs in the ALSPAC cohort. The ARIES participants were selected based on availability of DNA samples at two time points for the mother (antenatal and at follow-up when the offspring were adolescents) and three time points for the offspring (neonatal, childhood (age 7) and adolescence (age 17)). DNA methylation data for cord blood in the neonates and peripheral blood in the children at age 7 were included in this analysis. Written informed consent has been obtained for all ALSPAC participants. Ethical approval for the study was obtained from the ALSPAC Ethics and Law Committee and the Local Research Ethics Committees.

ALSPAC maternal smoking variables

Information on mothers' smoking status during pregnancy was obtained in questionnaires administered at 18 and 32 weeks gestation. Information was obtained about whether the mother smoked in each trimester of pregnancy and the number of cigarettes smoked on average per day. From these data, a dichotomous variable for sustained maternal smoking during pregnancy was derived. A mother was classified as a sustained smoker if she smoked in all three trimesters, smoked in the first and third trimester but not the second, or smoked in the second and third trimesters but not the first. The reference group consisted of mothers who had reported not smoking in all three trimesters. We excluded all individuals who smoked in one trimester only (i.e. not sustained) or who had missing information of smoking for

two or more trimesters. Of those with missing information on one trimester, women were classified as a sustained smoker if they said they smoked in the other two trimesters.

Another dichotomous variable for any maternal smoking during pregnancy was derived. Any smoking was defined based on reported smoking in any of the three trimesters. The reference group considered of mothers who had reported not smoking in all three trimesters.

ALSPAC methylation measurements

Cord blood and peripheral blood samples (whole blood, buffy coats or blood spots) were collected according to standard procedures. The DNA methylation wet-lab and pre-processing analyses were performed at the University of Bristol as part of the ARIES project. Following extraction, DNA was bisulphite-converted using the Zymo EZ DNA MethylationTM kit (Zymo, Irvine, CA). Following conversion, genome-wide methylation status of over 485,000 CpG sites was measured using the Infinium HM450 BeadChip according to the standard protocol. The arrays were scanned using an Illumina iScan and initial quality review was assessed using GenomeStudio (version 2011.1). Samples from all time points in ARIES were distributed across slides using a semi-random approach (sampling criteria were in place to ensure that all time points were represented on each array) to minimise the possibility of confounding by batch effects. In addition, during the data generation process, a wide range of batch variables were recorded in a purpose-built laboratory information management system (LIMS). The main batch variable was found to be the bisulphite conversion (BCD) plate number. Samples were converted in batches of 48 samples and each batch identified by a plate number. The LIMS also reported quality control (QC) metrics from the standard control probes on the 450K BeadChip for each sample. Samples failing QC (average probe p value >= 0.01) were repeated and if unsuccessful excluded from further analysis. As an additional QC step genotype probes were compared with SNP-chip data from the same individual to identify and remove any sample mismatches. For individuals with no genome-wide SNP data, samples were flagged if there was a sex-mismatch based on X-chromosome methylation.

For the secondary model, methylation data were pre-processed using R (version 3.0.1), with background correction and subset quantile normalization performed using the pipeline described by Touleimat and Tost.[3] All 485,512 probes were included in this analysis.

ALSPAC covariates

Maternal age at delivery was derived from date of birth, which was recorded at that time. This was categorized in 0-24, 25-29 or 30+ years. Mother's parity and her highest educational

qualification were recorded in a questionnaire completed during pregnancy. Parity was categorized into 0, 1, 2 or 3+ previous offspring. Maternal education was collapsed into one of four categories: vocational/CSE (the lower level of national school exams at age 16), O-level (the higher level of national school exams at age 16), A-level (national school exams at age 18) or university degree. Analyses were additionally adjusted for batch effects by adding bisulfite conversion (BC) run date as a covariate.

BAMSE

BAMSE design and study population

BAMSE is a prospective population-based cohort study of children recruited at birth and followed during childhood and adolescence. Details of the study design, inclusion criteria, enrollment and data collection are described elsewhere.[4] In short, 4,089 children born between 1994 and 1996 in four municipalities of Stockholm County, Sweden were enrolled. At baseline, when the infant was approximately 2 months of age, parents completed a questionnaire that assessed residential characteristics, as well as socioeconomic and lifestyle factors, including parental smoking habits. When children were 1, 2, 4, 8 years, the parents completed questionnaires focusing on children's symptoms related to wheezing and allergic diseases, as well as various exposures. The survey response rates were 96%, 94%, 91%, and 84%, respectively. Furthermore, blood was obtained from 2,614 (64%) and 2,480 (61%) of the children at the age of 4 and 8 years, respectively. The baseline and follow-up studies were approved by the Regional Ethical Review Board, Karolinska Institutet, Stockholm, Sweden, and the parents of all participating children provided informed consent.

BAMSE maternal smoking variables

Maternal smoking during pregnancy was assessed by questionnaire at the time of recruitment (at a median age of the children of 2 months). Sustained smoking in pregnancy was defined as maternal daily smoking of one cigarette or more during each trimester of pregnancy reported. Those smoking early in pregnancy and having missing information for later in pregnancy were classified as a sustained smoker. Those who smoked in trimester 1, did not smoke in trimester 2 but took it up again in trimester 3 were designated as a sustained smoker. Women who smoked in pregnancy but quit during the pregnancy or those started to smoke only in the 3rd trimester were excluded. Reference category included women who did not smoke

during pregnancy. Any smoking was defined as smoking at least one cigarette per day at any time during the pregnancy.

BAMSE methylation measurements

Epigenome-wide DNA methylation was measured in 472 Caucasian children, using DNA extracted from blood samples collected at the age of 8 years. 500 ng DNA per sample underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Shallow) (Zymo Research Corporation, Irvine, USA). Samples were plated onto 96-well plates in randomized order. Samples were processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA).

Quality control of analyzed samples was performed using standardized criteria. Samples were excluded in case of sample call rate <99%, colour balance >3, low staining efficiency, poor extension efficiency, poor hybridization performance, low stripping efficiency after extension and poor bisulfite conversion. We also applied multidimensional scaling (MDS) plot to evaluate gender outliers based on chromosome X data that produced two separated clusters for male and female. We omitted five samples that do not belong to the distinct cluster. Furthermore, we applied median intensity plot for methylated and unmethylated intensity by using the minfi R package (three samples below the 10.5 cutoff were excluded). All criteria described above led to exclusion of 8 samples. In addition, 89 samples were removed as these were already included in the MeDALL cohorts' sample, leaving a total of 375 samples in the analysis.

Probes with a single nucleotide polymorphism in the single base extension site with a frequency of > 5% were excluded,[5] as were probes with non-optimal binding (non-mapping or mapping multiple times to either the normal or the bisulphite-converted genome), and the probes belonging to chrX and chrY, resulting in the exclusion of 46,799 probes, leaving a total of 438,713 probes in the analysis.

Furthermore, we implemented "DASEN" recommended from watermelon package to do signal correction and normalization.[6]

BAMSE covariates

Information on parity was collected from the Birth Registry and was categorized into 0, 1 and ≥2. Maternal age and maternal education information were collected from a questionnaire administered at enrollment. Maternal age was used as a continuous covariate and maternal education was categorized into < high school, high school, some college, and ≥4 years of college. Ever doctor's diagnosis of asthma was considered to be a selection factor and

information was collected from questionnaires at 1, 2, 4, and 8 years of age in the children. Environmental tobacco smoke exposure at 8 years was evaluated by asking whether someone in the home smoked using questionnaires. The covariate batch was also accounted for in analysis models based on the bisulfite treatment date (6 categories).

CHAMACOS

CHAMACOS design and study population

The Center for the Health Assessment of Mothers and Children of Salinas (CHAMACOS) study is a longitudinal birth cohort study of the effects of exposure to pesticides and environmental chemicals on the health and development of Mexican-American children living in the agricultural region of Salinas Valley, CA. Detailed description of the CHAMACOS cohort has previously been published.[7,8] Briefly, 601 pregnant women were enrolled in 1999-2000 at community clinics and 527 liveborn singletons were born. Follow up visits occurred at regular intervals throughout childhood. Study protocols were approved by the University of California, Berkeley Committee for Protection of Human Subjects and written informed consent was obtained from all mothers.

CHAMACOS maternal smoking variables

Information on maternal smoking status was obtained through participant interview at baseline (~13 weeks gestation), follow up interview (~26 weeks gestation), and delivery. Subjects were considered sustained smokers if they indicated they had smoked since baseline during either follow up interview, or 'any' smokers if they reported having smoked at any of the three visits. However, if smoking was reported at baseline but not at follow up interview or delivery, participants were considered to have quit. Participants were designated as exposed to secondhand smoke if they weren't active smokers but reported living with a smoker during pregnancy or delivery interview.

CHAMACOS methylation measurements

DNA methylation was measured in DNA isolated from the cord blood of 378 CHAMACOS newborns by Illumina Infinium HumanMethylation450 (450K) BeadChips. DNA samples were bisulfite converted using Zymo Bisulfite Conversion Kits (Zymo Research, Irvine, CA), whole genome amplified, enzymatically fragmented, purified, and applied to the 450K BeadChips (Illumina, San Diego, CA) according to manufacturer protocol. 450K BeadChips

were handled by robotics and analyzed using the Illumina Hi-Scan system. DNA methylation was measured at 485,512 CpG sites.

Probe signal intensities were extracted by Illumina GenomeStudio software (version XXV2011.1, Methylation Module 1.9) methylation module and background subtracted. QA/QC was performed systematically by assessment of assay repeatability batch effects using 38 technical replicates, and data quality established as previously described.[9] Quality was also ensured by only retaining samples where 95% of sites assayed had detection P> 0.01. The same threshold (95% detection at p>0.01) was imposed to CpGs as well (n= 460 removed). Sites with annotated probe SNPs and with common SNPs (minor allele frequency >5%) within 50bp of the target identified in the MXL (Mexican ancestry in Los Angeles, California) HapMap population were excluded from analysis (n=49,748). This left a total of 435,369 CpGs in the analysis. In the secondary model, color channel bias, batch effects and difference in Infinium chemistry were minimized by application of ASMN algorithm,[9] followed by BMIQ normalization.[10]

CHAMACOS covariates

Maternal age, parity and education were assessed by participant interview at baseline visit (~13 weeks gestation). Maternal age was treated as a continuous variable. Parity was coded as a binary variable, with 0 as the baseline and ≥1 as the alternative. Maternal education was also treated as a binary categorical variable, with subjects either reporting less than a high school degree or having completed high school education or beyond. Analysis was also adjusted for batch effects by including 450K plate (n=10) as additional covariates.

CHS

CHS design and study population

The Children's Health Study (CHS) is a population-based prospective cohort study from age 5 onwards in Southern California, which has been described in detail elsewhere.[11] The study protocol was approved by the University of Southern California Institutional Review Board and informed, written consent and assent were provided by the parents and children respectively. A total of 5,341 children were recruited, all of whom were born between 1995 and 1997 and are currently being followed until age 18.

CHS maternal smoking variables

Assessment of prenatal tobacco smoke exposure was based on parent/guardian written responses on a self-administered questionnaire. Prenatal smoke exposure was defined as an affirmative answer to the following question: “Did your child’s biologic mother smoke while she was pregnant with your child (include time when she was pregnant but did not yet know that she was)?”

CHS methylation measurements

Epigenome-wide DNA methylation was measured in 85 Hispanic and non-Hispanic white children, using DNA extracted from newborn bloodspots archived by the state of California. Laboratory personnel performing DNA methylation analysis were blinded to study subject information. DNA was extracted from whole blood cells using the QiaAmp DNA blood kit (Qiagen Inc, Valencia, CA) and stored at -80 degrees Celcius. 700-1000ng of genomic DNA from each sample was treated with bisulfite using the EZ-96 DNA Methylation Kit™ (Zymo Research, Irvine, CA, USA), according to the manufacturer’s recommended protocol and eluted in 18 ul. The results of the Infinium HumanMethylation450 BeadChip (HM450) were compiled for each locus as previously described and were reported as beta (β) values.[12]

Quality control of analyzed samples was performed using standardized criteria. Samples were excluded in case of sample call rate <99% or mismatched sex, leading to exclusion of 3 subjects.

CpG loci on the HM450 array were removed from analyses if they were on the X and Y chromosomes, or if they contained SNPs, deletions, repeats, or if they have more than 10% missing values. Data were processed in the methylumi package in R, after which a normal exponential background correction was applied to the raw intensities at the array level to reduce background noise.[13] We then normalized each sample’s methylation values to have the same quantiles to address sample to sample variability.[14] Beta-values were calculated for all CpG sites.

CHS covariates

Information on maternal age, parity and maternal education was collected by questionnaire at enrollment. Maternal age was used as a continuous covariate. Parity was categorized into 0 or >=1 (*note, this collapsed categorical variable was to aid convergence issues for the cohort)*. Maternal education was categorized into lower (none, primary or secondary education) and higher (more than secondary education). Ancestry (European,

African, and Asian) was measured using ancestry informative markers (AIMS) SNPs and included as an additional covariate.

GALA II

GALA II design and study population

The Genes-environments & Admixture in Latino Americans (GALA II) study is a case-control study initiated in 2008 designed to investigate genetic, behavioral, social, and environmental determinants of asthma risk and morbidity among children aged 8-21 years, as previously described in detail.[15-17] The study used identical protocols to recruit nearly 5,000 Latinos (age 8-21) from 5 recruitment centers across the US (San Francisco Bay area; Houston, TX; Chicago, IL; New York, NY; and Puerto Rico). The study was approved by each of the five sites' institutional review boards, and all subjects provided informed consent/assent.

GALA II maternal smoking variables

Maternal smoking during pregnancy was assessed by self-report at time of subject recruitment. The child's parent or caregiver was asked, "Did [child's] mother smoke while she was pregnant with child?" Mothers who reported to have not smoked during pregnancy were considered the reference group of non-smokers. If the parent or caregiver responded affirmatively, a follow-up question was asked: "During which trimesters?" to which respondents could answer "Yes", "No", or "Don't know" to each of the three trimesters of pregnancy. Mothers who smoked for 1 or 2 trimesters were defined as having smoked during pregnancy but quit during pregnancy. Mothers who reported to have smoked for all three trimesters were considered sustained smokers.

GALA II methylation measurements

After examining DNA from 576 subjects for complete bisulfite conversion of DNA (Zymo Research, Irvine, CA), we randomized the samples onto the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA). Raw genome-wide methylation data were loaded in the R[18] package minfi[19] and assessed for basic quality control metrics, including determination of poorly performing probes with insignificant detection p values above background control probes (i.e., detection p value >0.01). Probes with a single nucleotide polymorphism in the single base extension site were excluded. Since our study

population included both males and females, we also removed the X and Y chromosomes from the raw methylation values. A total of 321,509 methylation loci were included for analysis.

We corrected for batch (microarray chip) effect using the ComBat function in the R package SVA (surrogate variable analysis)[20] and performed SWAN normalization to correct for intra-array differences between Illumina Type I and Type II probes.[21,22] A total of 569 samples passed quality control metrics and were included in the analysis.

GALA II covariates

Since GALA II was designed as a case-control study of asthma (i.e., subject enrollment was stratified by asthma status) of children and young adults age 8-21, we included asthma status, sex, and child's age as covariates in our secondary model. Because of concerns regarding postnatal secondhand smoke (SHS) exposure, we also adjusted for SHS exposure in the first two years of life (yes/no) as well as subjects' exposure to household smokers at time of recruitment (yes/no). These variables were chosen from postnatal SHS measure of other timepoints to maximize exposure assessment while minimizing multicollinearity. Additional covariates, also assessed by questionnaire, included maternal age (continuous), parity (categorical: 0, 1, 2, or ≥3 siblings), maternal education (some high school, high school graduate or equivalent, some college, and at least college graduate), and proportions of genetic African and Native American ancestry.

GECKO Drenthe

GECKO Drenthe design and study population

The GECKO Drenthe cohort is a population-based birth cohort in Drenthe, a northern province in the Netherlands.[23] All mothers of babies born between April 2006 and April 2007 were invited to participate during the third trimester of pregnancy. Of all 4,778 infants born in this period, a total of 2,874 children (60%) participated in the study and are followed until adulthood. This study has been approved by the Medical Ethical Committee of the University Medical Center Groningen and parents of all participants gave written informed consent.

GECKO maternal smoking variables

Information on maternal and paternal smoking was derived from the questionnaires during the third trimester of pregnancy. This questionnaire included a question about current smoking (during the third trimester) and a question whether she had ever smoked during this

pregnancy. If the mother currently smoked (in the third trimester), she was defined as a sustained smoker (n=70). If the mother had smoked during this pregnancy, but was not currently smoking anymore, the mother was defined as having smoked but quit by late in pregnancy (n=59). For the any smoking during pregnancy variable, we combined these two categories (n=129). If she had answered “no” to both questions, the mother was defined as a non-smoker (n=126). We selected 258 infants: 129 exposed to maternal smoking during pregnancy and 129 unexposed to both maternal and paternal smoking during pregnancy.

GECKO methylation measurements

From these 258 infants, we used DNA which was extracted from cord blood for the epigenome-wide DNA methylation analyses. To limit batch effects, we randomized all samples on gender and smoking status. Samples (500 ng per sample) were placed on three 96-well plates. Bisulfite conversion was performed using the EZ-96 DNA methylation kit (Zymo research Corporation, Irvine, USA). Then, we processed the samples with the Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA).

We used minfi to calculate betas and p values for all 485,577 CpGs. During the quality control, we excluded two males that clustered in the female group, based on X chromosome betas. We performed Illumina-suggested background normalization, colour correction and Subset-quantile Within Array Normalization (SWAN). We excluded one sample because it did not meet the criteria of ≥99% of the CpGs with detection p value<0.05. This resulted in 129 exposed and 126 unexposed children. We excluded control probes, probes on X or Y chromosomes and probes that did not meet our criteria of a detection p value of <0.05 in ≥99% of the samples, resulting in 465,891 remaining CpGs.

GECKO covariates

We adjusted analyses for maternal age, parity, maternal education and batch. Information on these covariates was collected with the questionnaire during the third trimester of pregnancy. Maternal age was calculated by subtracting the date of birth of the mother from the date of birth from the child, used as a continuous variable. Parity was categorized into 0, 1, 2 or 3+. Maternal education was categorized into low (everything lower than (applied university) versus high education (applied university). Covariate batch was categorized into 96-well plate number 1, 2 or 3.

Generation R

Generation R design and study population

The Generation R Study is a population-based prospective cohort study from fetal life onwards in Rotterdam, the Netherlands, which has been described in detail elsewhere.[24,25] The study protocol was approved by the Medical Ethics Committee of the Erasmus Medical Center, Rotterdam. Written informed consent was obtained for all participants. All children were born between April 2002 and January 2006 and form a largely prenatally enrolled birth cohort that is currently being followed until young adulthood. A total of 9,778 mothers were included, most during pregnancy (response rate at birth 61%).

Generation R maternal smoking variables

Maternal smoking during pregnancy was assessed by questionnaires in early (<18 weeks gestational age), mid (18-25 weeks gestational age) and late (>25 weeks gestational age) pregnancy. Pregnant women were asked whether they had smoked and if so, how much. Sustained smoking was defined as continued smoking of >=1 cigarette per day throughout pregnancy. Women who quit smoking during pregnancy were not classified as sustained smokers and were not included in the sustained smoking analysis. Any smoking was defined as smoking any number of cigarettes at any time during pregnancy.

Generation R methylation measurements

Epigenome-wide DNA methylation was measured in 979 Caucasian children, using DNA extracted from cord blood. 500 ng DNA per sample underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Shallow) (Zymo Research Corporation, Irvine, USA). Samples were plated onto 96-well plates in no specific order. Samples were processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA).

Quality control of analyzed samples was performed using standardized criteria. Samples were excluded in case of sample call rate <99%, colour balance >3, low staining efficiency, poor extension efficiency, poor hybridization performance, low stripping efficiency after extension and poor bisulfite conversion, leading to exclusion of 7 samples (6 for low sample call rates, 1 for poor bisulfite conversion). In addition, 2 samples were excluded because of a gender mismatch and 1 sample because of a retracted informed consent, leaving a total of 969 samples in the analysis.

Probes with a single nucleotide polymorphism in the single base extension site with a frequency of > 1% in the GoNLv4 reference panel were excluded, as were probes with non-

optimal binding (non-mapping or mapping multiple times to either the normal or the bisulphite-converted genome), resulting in the exclusion of 49,564 probes, leaving a total of 436,013 probes in the analysis.

Data were normalized with DASES normalization using a pipeline adapted from that developed by Touleimat and Tost.[3] DASES normalization includes background adjustment, between-array normalization applied to type I and type II probes separately, and dye bias correction applied to type I and type II probes separately and is based on the DASEN method described by Pidsley et al, but adds the dye bias correction, which is not included in DASEN.[6] Beta-values were calculated for all CpG sites.

Generation R covariates

Information on maternal age, parity and maternal education was collected by questionnaire at enrollment. Maternal age was used as a continuous covariate. Parity was categorized into 0 or >=1 (note, this collapsed categorical variable was to aid convergence issues for the cohort). Maternal education was categorized into lower (none, primary or secondary education) and higher (more than secondary education) (categories for maternal education were also collapsed to aid convergence issues). Analyses were additionally adjusted for batch effects by adding plate number (11 categories) as a covariate.

IOW

IOW design and study population

A whole population birth cohort was established on the Isle of Wight, UK, in 1989 to prospectively study the natural history of allergic diseases from birth onwards. Both the Isle of Wight and the study population are 99% Caucasian. Ethics approvals were obtained from the Isle of Wight Local Research Ethics Committee (now named the National Research Ethics Service, NRES Committee South Central – Southampton B) at recruitment and for the 1, 2, 4, 10 and 18 years follow-up. Of the 1,536 children born between January 1, 1989, and February 28, 1990, written informed consent was obtained from parents to enroll 1,456 newborns. Children were followed up at the ages of 1 (n = 1,167), 2 (n = 1,174), 4 (n = 1,218), 10 (n = 1,373), and 18 years (n = 1,313). From January 2012 to May 2014, we further recruited 367 1989-1990 cohort participants and 90 newborns of these participants. These 90 mother-child pairs are included in the current study.

IOW third generation cohort maternal smoking variables

Maternal smoking during pregnancy was assessed by questionnaires at 20 and 28 weeks of pregnancy, and 3 months after birth. Pregnant women were asked whether they had smoked and if so, how much. Sustained smoking was defined as continued smoking of ≥1 cigarette per day throughout pregnancy. Women who quit smoking during pregnancy were not classified as sustained smokers and were not included in the sustained smoking analysis. Any smoking was defined as smoking any number of cigarettes at any time during pregnancy.

IOW third generation cohort methylation measurements

Epigenome-wide DNA methylation was measured in 90 Caucasian children, using DNA extracted from cord blood, and 1000 ng DNA per sample underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Shallow) (Zymo Research Corporation, Irvine, USA). Samples were plated onto 96-well plates in no specific order. Samples were processed with the Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA).

Quality control of analyzed samples was performed using standardized criteria. All samples are over 99.8% probes detected. Samples were deleted if more than 75% CpG sites of that sample with detection p value larger than or equal to 10^{-5}, CpG sites were deleted if 10% of the samples with detection p value larger than or equal to 0.01. CpG sites with missing values caused by insufficient copies of a probe binding to the sample DNA were excluded from the study. Also, DNA methylation from the 90 subjects were measured in four batches. After excluding missing values in each batch, in total 358,214 CpG sites were included in the study.

Data were pre-processed using IMA package[26] in R including quantile normalization and type I and type II probe peak correction. Beta-values were calculated for all CpG sites.

IOW third generation cohort covariates

Information on maternal age, maternal education was collected by questionnaire at enrollment. Maternal age was used as a continuous covariate. Maternal education was categorized into four levels (<high school, High school, some college, and 4 years of college or more). Analyses were additionally adjusted for batch effects (4 batches) as a covariate.

MeDALL

MeDALL (Mechanisms of the Development of ALLergy) is a collaborative project supported by the European Union under the Health Cooperation Work Programme of the 7th

Framework programme (grant agreement number 261357). MeDALL epigenetics studies include four birth cohorts. These are EDEN, BAMSE, PIAMA and INMA.

MeDALL - INMA (Childhood and Environment)

The INMA—INfancia y Medio Ambiente—(Environment and Childhood) Project is a network of birth cohorts in Spain that aim to study the role of environmental pollutants in air, water and diet during pregnancy and early childhood in relation to child growth and development.[27] The study has been approved by Ethical Committee of each participating center and written consent was obtained from participating parents. Data for this study came from INMA Sabadell cohort (children born between 2004 and 2007).[27] A total of 203 mothers with pregnancy smoking information and offspring cord blood DNA methylation were included.

Blood at birth and at age 4y was obtained in EDTA tubes and extracted using the Chemagic DNA Blood Kits (Perkin Elmer) in a Chemagen Magnetic Separation Module 1 station at the Spanish National Genotyping Center (CEGEN, http://www.usc.es/cegen/). All additional laboratory methods related to DNA methylation measurements are described in the DNA methylation data section for the MeDALL consortium.

Maternal smoking during pregnancy was assessed by questionnaires at 32 weeks of pregnancy. Pregnant women were asked whether they were current smokers (at week 32 of pregnancy) and if so, how much. They were also asked if they had stopped smoking due to pregnancy and when (before pregnancy or at what month of pregnancy). Sustained smoking was defined as continued smoking of >=1 cigarette per day throughout pregnancy. Women who quit smoking during pregnancy were not classified as sustained smokers and were not included in the sustained smoking analysis. Any smoking was defined as smoking any number of cigarettes at any time during pregnancy.

Information on parity and maternal education was collected by questionnaire at enrollment (week 12 of pregnancy). Maternal age was used as a continuous covariate. Parity was categorized into 0 or >=1. Maternal education was categorized into three levels: primary or less, medium (secondary) or high (university).

MeDALL – INMA Gene expression data

At age four years, whole blood was collected in PAXGene tubes and extracted using the kit recommended by the company. All samples had a RNA Integrity Number (RIN) higher than seven. Gene expression data was obtained using Affymetrix HTA 2.0 at the European Institute for Systems Biology and Medicine in Lyon. Gene expression was normalized using Expression

Console Software from Affymetrix and probes were clustered to the transcript level. Twelve samples were excluded because they were outliers defined as more than 3SD from the mean for PC1 or PC2 (N=8) or there were sex discrepancies (N=8). Expression transcripts were annotated using version 35 of Affymetrix annotation. The final sample size was 107.

MeDALL - EDEN

The EDEN (Etude des Déterminants pré et post natals du développement et de la santé de l'Enfant) study is a prospective Birth Cohort Study (https://eden.vjf.inserm.fr/), which has been described in detail elsewhere.[28] Pregnant women seen for a prenatal visit at the departments of Obstetrics and Gynecology of the University Hospitals of Nancy and Poitiers before their twenty-fourth week of amenorrhea were invited to participate. Enrollment started in February 2003 in Poitiers and September 2003 in Nancy; recruitment lasted 27 months in each center. Among eligible women, 55% (2,002 women) accepted to participate. The study has been approved by the ethical committees Comité Consultatif pour la Protection des Personnes dans la Recherche Biomédicale, Le Kremlin-Bicêtre University hospital, and Commission Nationale de l'Informatique et des Libertés .

Maternal smoking during pregnancy was assessed by questionnaires completed in mid-pregnancy (24-28 weeks gestational age) and after delivery. Pregnant women were asked whether they had smoked and if so, how much. Sustained smoking was defined as continued smoking of ≥1 cigarette per day throughout pregnancy. Women who quit smoking during pregnancy were not classified as sustained smokers and were not included in the sustained smoking analysis. Any smoking was defined as smoking any number of cigarettes at any time during pregnancy.

MeDALL - BAMSE

See description of the BAMSE study earlier in the study population methods. The BAMSE MeDALL subset consisted of 289 samples collected at 4 years processed as described for the MeDALL consortium (see MeDALL methylation measurements section).

MeDALL - PIAMA

For the PIAMA (Prevention and Incidence of Asthma and Mite Allergy) birth cohort study, pregnant women were recruited in 1996-1997 during their second trimester of pregnancy from a series of communities in the North, West, and Centre of The Netherlands. Details of the study design have been published previously.[29] Non-allergic pregnant women were invited to

participate in a “natural history” study arm. Pregnant women identified as allergic through a validated screening questionnaire were primarily allocated to an intervention arm with a random subset allocated to the natural history arm. The intervention involved the use of mite-impermeable mattress and pillow covers. The study started with 3,963 newborns. Full questionnaire follow-ups of the children took place at 3 months of age, yearly from 1 to 8 years of age, and at ages 11 and 14 years. Medical examinations were performed in subsets of the population at ages 4, 8, 12 and 16 years. DNA was extracted of children who provided blood samples at ages 4 and 8 years. The Medical Ethical Committees of the participating institutes approved the study, and all participants gave written informed consent. Any maternal smoking during pregnancy was defined as smoking any number of cigarettes at least during the first 4 weeks of pregnancy. Sustained maternal smoking during pregnancy was defined as smoking any number of cigarettes during the third trimester of pregnancy. Childhood smoke exposure was defined as any smoking by the father or the mother inside the house between the child’s birth and 8 years of age.

MeDALL - Methylation measurements

In the MeDALL study, peripheral blood samples were collected from all consenting cohort participants, and DNA from peripheral and cord blood samples was isolated by the laboratories participating in the MEDALL study using different methods. To uniform the concentration and purity the samples underwent a precipitation-based concentration and purification using GlycoBlue (Ambion) if needed. DNA concentration was determined by Nanodrop measurement and picogreen quantification. After normalization of the concentration, the samples were randomized to avoid batch effects. Standard male and female DNA samples were included in this step for control reasons. 500 ng of DNA of each sample was bisulfite-converted using the EZ 96-DNA methylation kit following the manufacturer’s standard protocol. After verification of the bisulfite conversion using Sanger Sequencing, the DNA methylation was measured using the Illumina Infinium HumanMethylation450 BeadChip. DNA methylation data were preprocessed using the minfi package,[19] and the DASEN method from the wateRmelon package[6] was used for normalization.

A series of steps were completed for quality control and data analysis. First, we implemented sample filtering to remove bad quality and mixed up samples. Second, 65 SNPs assays, the probes on sex chromosomes, the probes that mapped on multi-loci, and the probes containing SNPs at the target CpG sites with a MAF>10% were excluded. The multi-loci probes and probes containing SNPs are selected based on reference.[5] This led to a total number of

439,306 CpG sites. Third, we implemented "DASEN" to perform signal correction and normalization. Fourth, to remove bias in methylation profiles unrelated to underlying biological processes, we implemented correction procedures based on 613 negative control probes presented in 450K arrays since these negative control probes are supposed to not relate to biological variation.[30] Finally, we implemented PCA on control probes data. We performed 10,000 permutations for controls probe data and selected principal components with p value defined as to get the p value of (number of var(random pc) > var(pc))/ (number of permutations) < 10^{-4}.[31] The methylation data for each CpG are thus the residuals from a linear model incorporating the significant 5 PCs. The final robust linear regression models were adjusted for maternal age, parity, maternal education, bclot, and sections, for both newborn and older children analyses. Here "bclot" represents bisulfite conversion kit batch number. "Sections" denotes the position of array. In newborns, there are only two cohorts (EDEN and INMA), and in older kids, there are four cohorts (EDEN, INMA, PIAMA and BAMSE), with an average age around 4. Parity was categorized into 0 and >=1 due to convergence issues using robust linear regression. The other covariates were selected based on testing the difference of the correlation p value before and after correction by Kolmogorov–Smirnov test.

MOBA

MoBa design and study population

Participants in the current analysis represent three subsets of mother-offspring pairs from the national Norwegian Mother and Child Cohort Study (MoBa).[32,33] Each subset is referred to here as MoBa1, MoBa2, and MoBa3. MoBa1 and MoBa2 study populations were part of a larger study within MoBa that was designed to evaluate the association between maternal plasma folate during pregnancy and childhood asthma status at 3 years of age.[34] We previously reported an association between maternal smoking during pregnancy and differential DNA methylation in 1,062 MoBa1 newborns.[35] We subsequently measured DNA methylation in an additional 685 newborns with maternal plasma folate measurements and following separate quality control and preprocessing (MoBa2). MoBa3 was designed to evaluate the association between differential cord blood DNA methylation and later childhood cancer status. These analyses include the children who had cord blood DNA methylation measurements, smoking information, and covariate data (N=1,063 from MoBa1; N=671 from MoBa2; N=207 from MoBa3), and each dataset was analyzed independently. The year of birth for participants in these MoBa participants ranged from 2000-2009. All three studies were approved by the

Regional Committee for Ethics in Medical Research and the Norwegian Data Inspectorate, and written informed consent was provided by all mothers participating. In addition, MoBa1 and MoBa2 were approved by the Institutional Review Board of the National Institute of Environmental Health Sciences, USA.

MoBa maternal smoking variables

For MoBa1 and MoBa2, maternal blood sample collection during pregnancy was completed as previously described.[33] Maternal blood samples were drawn during pregnancy (median weeks gestation=18 weeks, 25^{th}-75^{th} percentile=16-21 weeks) in EDTA-coated tubes, centrifuged within 30 minutes after collection, and stored at 4°C in the hospital where they were collected. Samples were then shipped overnight to the Biobank of MoBa at the Norwegian Institute of Public Health in Oslo. Upon receipt (1-2 days after blood collection), plasma was aliquoted onto polypropylene microtiter plates, sealed with heat-sealing foil sheets, and stored at -80°C. Maternal blood samples were not analyzed for MoBa3 for this analysis.

Maternal smoking during pregnancy was assessed by maternal questionnaire for all three datasets. For all MoBa1 subjects, 221 of the MoBa2 subjects, and no MoBa3 subjects, cotinine, a biomarker of smoking, was measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) [36] in plasma collected at approximately gestational week 18. Cotinine values above 56.8 nmol/L were used to indicate that a mother was smoking at this time point.[37] The self-reported smoking information and cotinine data were combined for samples with both information as described in a previous publication using the MoBa1 dataset.[38] For this analysis, we classified mother's smoking into three categories using her report of smoking during pregnancy and cotinine values: never smoking during pregnancy, smoked during pregnancy but quit by 18 weeks, and smoked through gestational week 18. Quitting by 18 weeks was defined by mother's report plus having a cotinine value below 56.8 nmol/L, if available. We further collapsed this variable to represent sustained smoking (yes/no) during pregnancy, excluding mothers who stopped smoking during pregnancy from the analyses to reduce noise. We also created a variable capturing any smoking (yes/no) during pregnancy, where mothers who stopped smoking during pregnancy were assigned a value of "yes" for any smoking during pregnancy.

MoBa covariates

For all MoBa datasets in this analysis, information on maternal age, parity, and maternal education was collected via questionnaires completed by the mother or from birth registry

records. Maternal age was included as a continuous variable. Parity was categorized as 0, 1, 2, or ≥ 3 births. Maternal educational level was categorized as previously described[35] into less than high school/secondary school, high school/secondary school completion, some college or university, or 4 years of college/university or more.

MoBa methylation measurements

Details of the DNA methylation measurements and quality control for the MoBa1 participants were previously described[35] and the same protocol was implemented for the MoBa2 participants. Briefly, umbilical cord blood samples were collected and frozen at birth at -80°C. All biological material was obtained from the Biobank of the MoBa study.[33] Bisulfite conversion was performed using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, CA) and DNA methylation was measured at 485,577 CpGs in cord blood using Illumina's Infinium HumanMethylation450 BeadChip.[39] Raw intensity (.idat) files were handled in R using the *minfi* package[19] to calculate the methylation level at each CpG as the beta-value (β=intensity of the methylated allele (M)/(intensity of the unmethylated allele (U) + intensity of the methylated allele (M) + 100)) and the data was exported for quality control and processing.

Probe and sample-specific quality control was performed in the MoBa1, MoBa2, and MoBa3 datasets separately. Similar protocols were applied to MoBa1 and Moba2, as follows: Control probes (N=65) and probes on X (N=11,230) and Y (N=416) chromosomes were excluded in both datasets. Remaining CpGs missing > 10% of methylation data were also removed (N=20 in MoBa1, none in MoBa2). Samples indicated by Illumina to have failed or have an average detection p value across all probes < 0.05 (N=49 MoBa1, N=35 MoBa2) and samples with gender mismatch (N=13 MoBa1, N=8 MoBa2) were also removed. For MoBa1 and MoBa2, we accounted for the two different probe designs by applying the intra-array normalization strategy Beta Mixture Quantile dilation (BMIQ).[10] The Empirical Bayes method via *ComBat* was applied separately in each dataset for batch correction using the *sva* package in *R.*[40]

Similar data quality control and processing was applied to MoBa3 with some slight differences. Methylation features were filtered from (i) cross-reactive probes, (ii) probes mapping to sex chromosomes and (iii) probes overlapping with a known single nucleotide polymorphism (SNP) with an allele frequency of at least 5% in the overall population (all ethnic groups), resulting in the exclusion of 36,231 probes. Data quality was further assessed using box plots for the distribution of methylated and unmethylated signals, and multidimensional scaling plots and unsupervised clustering were used to check for sample outliers. After

background correction and color-bias adjustment, type I and type II probe distributions were aligned using the intra-array beta-mixture quantile normalization (BMIQ)[10] from the wateRmelon package. Beta-values were calculated for all CpG sites. Batch effects were corrected by surrogate variable analysis (SVA).[20]

NEST

NEST design and study population

The Newborn Epigenetics STudy (NEST) is a multiethnic birth cohort designed to identify the effects of early exposures on epigenetic profiles and phenotypic outcomes. Pregnant women were recruited from prenatal clinics serving Duke University Hospital and Durham Regional Hospital Obstetrics facilities in Durham, North Carolina from April 2005 to July 2009. Gestational age at enrollment ranged from 6 to 42 weeks (median 30 weeks). Eligibility criteria were women aged 18 years or older, English speaking, pregnant, and an intention to use one of the two obstetrics facilities. Among these, women infected with HIV or intending to give up custody of the offspring of index pregnancy were excluded. Current smokers were targeted for the first ~200 participants. Of the 1,101 women who met eligibility criteria and were approached, 895 (81%) were enrolled and umbilical cord blood was collected from 741 infants. The current analysis was limited to the 413 infants with 450k and covariate data. This study was approved by the Duke Institutional Review Board.

NEST maternal smoking variables

Four questions were used to ascertain smoking status. Women were first asked if they ever smoked and whether they were current smokers by responding to questions “Have you ever smoked 100 cigarettes or more in your lifetime?” (Yes/ No), followed by, “Do you smoke now?” (Yes/No). To determine the timing of cigarette smoking exposure to the offspring, women who reported being smokers were then asked to respond to the question, “Did you smoke anytime in the year before you found out you were pregnant?” Women also responded to the question “After you found out you were pregnant, which of the following best describes your behavior?” The four possible responses were, “I continue to smoke,” or “I stopped during the first/second/third trimester.” From these responses, three categories of maternal cigarette smoking were created as follows: (a) “smokers during pregnancy” were women who reported having ever smoked 100 cigarettes or more, being a current smoker, and smoking during pregnancy; (b) “quitters during pregnancy” were women who reported having ever smoked 100

cigarettes or more, smoking during the year of pregnancy, and stopping smoking any time during the pregnancy; (c) “non-smokers” were women who never smoked 100 cigarettes or more.

NEST methylation measurements

Genomic DNA from buffy coat specimens was extracted from umbilical cord blood using Puregene Reagents (Qiagen, Valencia, CA). Bisulfite conversion was performed using the EZ-96 DNA Methylation Kit (Zymo Research Corporation) and DNA methylation was measured at 485,577 CpGs using Illumina’s Infinium Human- Methylation450 BeadChip. Illumina’s GenomeStudio Methylation module version 1.0 (Illumina Inc.) was used to calculate the methylation level at each CpG as the beta value.

Probe and sample-specific quality control was performed in the NEST cohort using a similar approach to MoBa1 and MoBa2 cohorts as the data analysis was completed at the NIEHS. Specifically, control probes (N=65) and probes on X (N=11,230) and Y (N=416) chromosomes were excluded as well as CpGs missing > 10% of methylation data. Samples indicated by Illumina to have failed or have an average detection p value across all probes < 0.05 and samples with gender mismatch were also removed. The two different probe designs by applying the intra-array normalization strategy Beta Mixture Quantile dilation (BMIQ).[10] The Empirical Bayes method via *ComBat* was applied for batch correction using the *sva* package in *R.*[40]

NEST covariates

Covariates considered as potential confounders were maternal body mass index (BMI) before pregnancy, age (<30 years, 30–39 years, >40+ years), educational attainment, and ethnicity (African American, Caucasia and Other), as well as offspring gestational age at delivery (<37 weeks or ≥37 weeks), birth weight, and sex.

NFCS

Norway Facial Clefts Study design and study population

The Norway Facial Clefts Study (NFCS) is a national population-based case-control study of cleft lip and cleft palate, disorders characterized by the incomplete fusion of the lip and/or palate during development. The study design has been previously described in detail.[41] Study approval was obtained by the Norwegian Data Inspectorate and Regional Medical Ethics

Committee of Western Norway and informed consent was provided by both the mother and father. Briefly, between the years of 1996 and 2001 all families of newborns referred for cleft surgery in Norway were contacted and, of those eligible, 88% agreed to participate (N=573). Controls were selected by a random sampling of roughly 4 per 1,000 live births in Norway during that same time period and, of those eligible, 76% agreed to participate (N=763). After completion of data collection and linkage with the Medical Birth Registry, all identifiers were permanently stripped from the data set, with no opportunity for further follow-up.

Norway Facial Clefts Study smoking variables

Information about maternal tobacco smoke exposure at the beginning of pregnancy was obtained through self-administered questionnaires sent to the mothers 3-4 months after delivery. Mothers were asked about cigarette smoking during the first trimester (average number of cigarettes smoked per day or per month). If a mother reported smoking at least one cigarette per day on average, she was considered to be an active smoker at the beginning of pregnancy. In order to determine if a woman was a sustained smoker (i.e. smoked at the beginning of pregnancy and did not report quitting at any time during pregnancy), we utilized the Medical Birth Registry of Norway (MBRN). The MBRN provides information on whether the mother smoked at the end of pregnancy, as well as the frequency of smoking. A woman was considered to be a sustained smoker if she reported actively smoking at the beginning of pregnancy based on the Norway Facial Clefts Study questionnaire and met one of the following criteria: 1) reported being a daily smoker at the end of pregnancy, 2) reported being a sometimes smoker at the end of pregnancy but indicated smoking at least one cigarette per day on average, or 3) no information regarding smoking status at the end of pregnancy was available (i.e. in the absence of further information, women were assumed to have continued smoking).

Norway Facial Clefts Study measurements

Epigenome-wide DNA methylation was measured in 889 newborns, using DNA extracted from heel stick blood samples that were collected 2-3 days after delivery as part of a standardized program of testing for phenylketonuria (PKU). A detailed description of DNA methylation data generation (Illumina HumanMethylation450 beadchips), quality control, and data pre-processing has been provided previously.[42] Briefly, one microgram of DNA was bisulfite converted using the EZ DNA Methylation kit following the manufacturer's protocol. 898 newborn and 60 technical control samples were run on Illumina HumanMethylation450

BeadChips according to the manufacturer's instructions at the NIH Center for Inherited Disease Research. After exclusions, 889 samples remained for analysis.

Raw intensity data were obtained using the Illumina GenomeStudio methylation module (version 2011.1). At each CpG site on the array, methylation status was determined based on intensity measures corresponding to unmethylated (U) or methylated (M) signal. The Illumina HumanMethylation450 BeadChip contains two probe types: Infinium Type I (2 probe types, 1 color channel) and II (1 probe type, 2 color channels). As the Type II probes use two color channels to assess methylation, dye bias was corrected using the normalization function (normalizeMethyLumiSet), provided in the R package, methylumi.[43] Before association analysis, the M and U intensity values for Type I and II probes were separately background adjusted (4 separate groups) using the robust multi-array average (RMA) method (Irizarry et al. 2003) and quantile normalized using the normalization function (normalize.quantiles), provided in the R package, Affy.[14] The β-value (M/(M+U+100)) was then computed and used in the association analysis. β-values that were more than 3 standard deviations from the mean and methylation levels that were deemed undetectable (Illumina detection p value ≥ 0.05) were excluded. Prior to running the robust linear regression secondary model, residuals were calculated for each CpG probe from linear regression models adjusting for batch (96-well plate, 10 levels). The mean β-value was added back to the residuals.

Norway Facial Clefts Study covariates

Information on maternal age, parity, and maternal education was collected in the same self-administered questionnaires (around 3-4 months after delivery). Facial cleft status was categorized as none (control), cleft lip with or without cleft palate, and cleft palate only. Maternal age was treated as a continuous covariate. Parity was categorized as 0, 1, 2, 3+. Maternal education was categorized as less than high school and high school and above. Analyses were additionally adjusted for bisulfite conversion efficiency by adding the mean bisulfite control probe intensity as a covariate in the model.

Rotterdam Study

Rotterdam study design and study population

The analyses for DNA methylation and gene expression were performed using data from the third cohort of the Rotterdam Study. The design of the Rotterdam Study has been described elsewhere.[44] In brief, all inhabitants living in the neighborhood Ommoord in Rotterdam, the

Netherlands, aged 45 years and over were invited to participate. During the center visit, 3,934 participants were examined between February 2006 and December 2008. We performed the analyses on 747 Caucasian subjects with DNA methylation data and gene expression data available. The Rotterdam Study has been approved by the medical ethics committee according to the Population Screening Act: Rotterdam Study, executed by the Ministry of Health, Welfare and Sports of the Netherlands. All participants in the present analysis provided written informed consent to participate and to obtain information from their treating physicians.

Rotterdam study DNA methylation data

DNA methylation was measured using Illumina Human Methylation 450K array of whole blood samples, following the manufacturers' protocol.[45] The methylation percentage of a CpG site was reported as a beta-value ranging between 0 (no methylation) and 1 (full methylation).

Quality control of the samples was done with Genome Studio and MixupMapper.[46] A total number of 16 samples were removed: 7 had a sample call rate below 99%; 5 had incomplete bisulfite conversion and 4 had gender clustering. Quality control of the probes was done based on the detection p value calculated with Genome Studio. Probes with a detection p value of more than 0.01 in more than 1% of the samples were excluded. This resulted in a total set of 474,528 probes which were normalized with Dasen. Dasen normalization involved background adjustment of the methylated and unmethylated intensities followed by separate quantile normalization of methylated Type I, unmethylated Type I, methylated Type II and unmethylated Type II intensities.[6]

Rotterdam study gene expression data

Whole-blood was collected (PAXGene Tubes – Becton Dickinson) and total RNA was isolated (PAXGene Blood RNA kits - Qiagen). To ensure a constant high quality of the RNA preparations, all RNA samples were analyzed using the Labchip GX (Calliper) according to the manufacturer's instructions. Samples with an RNA Quality Score more than 7 were amplified and labeled (Ambion TotalPrep RNA), and hybridized to the Illumina HumanHT12v4 Expression Beadchips as described by the manufacturer's protocol. Processing of the Rotterdam Study RNA samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Centre Rotterdam. The accession number for the RS-III expression dataset reported in this paper is GEO: GSE33828. Illumina gene expression data was quantile-normalized to the median distribution and subsequently log2-transformed. The probe and sample means were centered to zero. Genes were declared significantly expressed when the

detection p values calculated by GenomeStudio were less than 0.05 in more than 10% of all discovery samples, which added to a total number of 21,238 probes.[46] Quality control was done using the eQTL-mapping pipeline. We only analyzed probes that uniquely mapped to the human genome build 37.[47]

SEED

SEED design and study population

The Study to Explore Early Development (SEED) is a multi-site US-based case-control study of autism that has been described in detail.[48] SEED phase I enrolled families with a child born between September 2003 and August 2006. All children enrolled in SEED I were aged 30-68 months at the time of clinical assessments and sample collection. IRB approval was obtained from each of six SEED study sites including Northern California Kaiser Permanente (CA), Johns Hopkins University (MD), University of North Carolina (NC), University of Pennsylvania (PA), University of Colorado Denver (CO), and the Centers for Disease Control (GA).

SEED maternal smoking variables

Maternal smoking during pregnancy was obtained via the SEED caregiver interview (CGI), as described previously [48]. In this analysis, we considered only maternal interviews. The SEED CGI instrument includes several questions regarding active smoking and dose (cigarettes/day) during pregnancy by month or trimester (for those not being able to recall consumption by month). Maternal active smoking during each trimester was defined as either any exposure for >2 months, or an average consumption of ≥1 cigarette/day for ≥1 month during the exposure window. We defined “sustained smokers” as having active maternal smoking for at least 2 of the 3 trimesters. Participants who were defined as active smokers during at least one trimester were considered “any smokers”.

SEED methylation measurements

Whole blood genomic DNA was extracted from 610 SEED samples and 500ng was bisulfite treated using the EZ DNA methylation kit (Zymo Research, Irvine, CA). Samples were randomized across and within plates and run on the Illumina HumanMethylation 450k array (Illumina, San Diego, CA) at the Center for Epigenetics, Johns Hopkins University. Each plate contained replicate samples, as well as two internal control samples used by the Epigenetics Center for cross-plate comparisons and quality control measures.

Quality assurance analyses were performed using Bioconductor and R-3.0.x. Illumina idat files were obtained and processed using the minfi package (version 1.8.9).[19] We generated sample quality control reports using the qcReport function. We assessed the correlation of replicate samples across plates to identify problems with particular plates/batches and to assess the accuracy of the DNA methylation values; correlation coefficients for the 8 replicate samples ranged from 0.989 to 0.994. Based on insufficient probe intensity in >10% of samples, 771 probes were excluded. Two samples with low overall 450K intensities and one sample with an outlier blood cell composition were removed, resulting in high quality DNAm data for 607 samples, 584 of those had both prenatal tobacco exposure and covariate data. We then performed quantile normalization[19] and adjusted normalized data for batch effects using the sva package (version 3.9.1).[49]

SEED covariates

Maternal age and education were obtained via the SEED CGI. Education was categorized as less than high school, high school, some college, college degree or more. Age was continuous. Parity was not available at the time of analysis. Our analyses were adjusted for autism status (yes, no). Analyses were also adjusted for ancestry groups (European, African, Asian, Admix), determined using corresponding GWAS data.

Project Viva

Project Viva design and study population

Project Viva is a population-based prospective pre-birth cohort of mothers and their children in Eastern Massachusetts, USA, which has been described in detail elsewhere.[50] The study has been approved by the Institutional Review Board of Harvard Pilgrim Health Care and written consent was obtained from participating women. Women were enrolled from 1999 to 2002 and enrollment included a total of 2,128 live births. Follow up of the children through early adolescence is ongoing.

Project Viva maternal smoking variables

Maternal smoking during pregnancy was assessed on questionnaires administered in early pregnancy asking about current smoking (mean 11.3 weeks), and about smoking in the past 3 months at the 2nd trimester visit and in an interview at delivery. Any smoking during pregnancy was defined as report of current smoking on the early pregnancy questionnaire or

smoking in the past 3 months on the mid-pregnancy or delivery questionnaires or extraction of maternal smoking from the medical record chart review.

Project Viva methylation measurements

DNA samples extracted from cord blood were arranged using a stratified randomization to ensure balance of cohort characteristics across sample plates/batches. Samples were bisulfite converted using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, USA). The Illumina Infinium HumanMethylation450 BeadChip (Illumina Inc., San Diego, USA) was run at Illumina FastTrack Microarray Services (San Diego, CA). Failing samples were rerun and passing arrays were defined as having >99% of probes with a detection p value <0.05. After excluding samples with identity concerns (inconsistent genotyping and/or inferred sex) there were 485 unique samples and all of these were included in the analysis of any smoking during pregnancy. Sample preprocessing included the exclusion of allosomal probes, non-CpG probes, and failing probes (<99% of samples with detection p values <0.05). A total of 14,707 probes were excluded from chromosomes X or Y or being a CpH site or an rs probe. An additional 5,918 probes were excluded for failing by the detection p value criteria above and results were reported for the remaining 464,952 CpG sites. Retained data with a detection p value > 0.05 was set to NA (but included no more than 5 samples per probe by the failure criteria described). Pre-processing for the secondary model included four additional steps: 1. background subtraction using the out-of-band probes (noob) as implemented in methylumi; 2. dye bias adjustment using the methylumi default; 3. within array type II probe adjustment using BMIQ as implemented in wateRmelon; and 4. conversion to M-values, adjustment for analytic plate using ComBat (plate was the largest batch effect seen in PCA), and subsequent transformation back to the beta scale.

Project Viva covariates

Information on maternal age, education, race/ethnicity, and parity was collected by interview or questionnaire at enrollment in the 1st trimester of pregnancy. Maternal age was used as a continuous covariate. Maternal education was categorized into lower (none, primary or secondary education) and higher (more than secondary education). Race/ethnicity of the mother was dichotomized as White/non-White. Parity was categorized into 0 or ≥1.

Supplemental Acknowledgements

ALSPAC: We are extremely grateful to all the families who took part in this study, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, and nurses. We would like to acknowledge Oliver Lyttleton, Sue Ring, Nabila Kazmi, and Geoff Woodward for their earlier contribution to the generation of ARIES data (ALSPAC methylation data).

BAMSE: We would like to thank all the families for their participation in the BAMSE study. In addition, we would like to thank Eva Hallner, Sara Nilsson, and André Lauber at the BAMSE secretariat for invaluable support, as well as Mutation Analysis Facility (MAF) at Karolinska Institutet for genome-wide methylation analysis, and Ingrid Delin for excellent technical assistance. The computations were performed on resources provided by SNIC through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) under Project b2014110.

CHAMACOS: We are grateful to the CHAMACOS staff, students, community partners, participants, and their families. We would also like to thank Dr. Kim Harley and Ms. Katherine Kogut for their contributions to this study.

CHS: We would like to express our sincere gratitude to Martin Kharrazi, Steve Graham, and Robin Cooley at the California Biobank Program and Genetic Disease Screening Program within the California Department of Public Health for their assistance and advice regarding newborn bloodspots. We are indebted to the school principals, teachers, students and parents in each of the study communities for their cooperation and especially to the members of the health testing field team for their efforts.

EDEN: The analysis for EDEN is the result of a "Collaboration INSERM et CEA-IG-CNG Epigenetique. On behalf of the EDEN Mother-Child Cohort Study Group, we thank the study participants and staff for their participation in this cohort.

GALA II: We thank the families, patients, and the numerous health care providers and community clinics for their support and participation in GALA II.

GECKO: We are grateful to the families who took part in GECKO, the midwives, nurses and GPs for their help in recruiting the data, and the whole team from GECKO.

Generation R: The Generation R Study is conducted by the Erasmus Medical Center in close collaboration with the School of Law and Faculty of Social Sciences of the Erasmus University Rotterdam, the Municipal Health Service Rotterdam area, Rotterdam, the Rotterdam Homecare Foundation, Rotterdam and the Stichting Trombosedienst & Artsenlaboratorium

Rijnmond (STAR-MDC), Rotterdam. We gratefully acknowledge the contribution of children and parents, general practitioners, hospitals, midwives, and pharmacies in Rotterdam. The generation and management of the Illumina 450K methylation array data (EWAS data) for the Generation R Study was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, the Netherlands. We thank Ms. Sarah Higgins, Ms. Mila Jhamai, Ms. Marjolein Peters, Dr. Lisette Stolk, Mr. Michael Verbiest, and Mr. Marijn Verkerk for their help in creating the EWAS database and the analysis pipeline.

INMA: INMA researchers would like to thank all the participants for their generous collaboration. INMA researchers are grateful to Silvia Fochs, Nuria Pey, and Muriel Ferrer for their assistance in contacting the families and administering the questionnaires. A full roster of the INMA Project Investigators can be found at http://www.proyectoinma.org/presentacion-inma/listado-investigadores/en_listado-investigadores.html.

IOW: IOW cohort acknowledges the great help from the nurses at the David Hide Asthma and Allergy Research Centre led by Professor Hasan Arshad, and Faisal Rezwan from the University of Southampton in DNA methylation data pre-processing. We greatly appreciate the participating families in the third generation study.

MoBa 1 and 2: The MoBa cohort acknowledges Shuangshuang Dai of Integrative Bioinformatics at the NIEHS and Jianping Jin of Westat for their programming assistance. We are also grateful to all the participating families in Norway who take part in the ongoing MoBa cohort study.

MoBa 3: We thank Dr. Florence Le Calvez-Kelm and Mr. Geoffroy Durand for helping in the HM450 experiments at IARC.

NFCS: We thank all individuals for participating in the Norway Facial Clefts Study.

PIAMA: The authors thank all the children and their parents for their cooperation. The authors also thank all the field workers and laboratory personnel involved for their efforts, and Marjan Tewis for data management.

Rotterdam Study: The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists in Rotterdam.

MeDALL: The authors warmly thank Camille Ménard and Stéphane Joly (European Institute for Systems Biology and Medicine, EISBM)for their tremendous and efficient work and involvement in the production of MeDALL gene expression data (BAMSE and INMA cohorts).

Funding Support

ALSPAC: The ALSPAC cohort was supported by the UK Medical Research Council Integrative Epidemiology Unit and the University of Bristol (MC_UU_12013_1, MC_UU_12013_2, MC_UU_12013_5 and MC_UU_12013_8), the Wellcome Trust (WT088806) and the United States National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK10324). RCR is funded by a Wellcome Trust 4-year PhD studentship (Grant Code: WT097097MF and 099873/Z/12/Z). GDS and CLR are partially supported by the ESRC (RES-060-23-0011) "The biosocial archive: transforming lifecourse social research through the incorporation of epigenetic measures" The UK Medical Research Council and the Wellcome Trust (Grant ref: 102215/2/13/2) and the University of Bristol provide core support for ALSPAC. The Accessible Resource for Integrated Epigenomics Studies (ARIES) was funded by the UK Biotechnology and Biological Sciences Research Council (BB/I025751/1 and BB/I025263/1). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

BAMSE: The BAMSE cohort was supported by The Swedish Research Council, The Swedish Heart-Lung Foundation, Freemason Child House Foundation in Stockholm, MeDALL (Mechanisms of the Development of ALLergy), a collaborative project conducted within the European Union (grant agreement No. 261357), Centre for Allergy Research, Stockholm County Council (ALF), Swedish foundation for strategic research (SSF, RBc08-0027, EpiGene project), the Strategic Research Programme (SFO) in Epidemiology at Karolinska Institutet, The Swedish Research Council Formas and the Swedish Environment Protection Agency.

CHAMACOS: The CHAMACOS study was supported by the NIH grants P01 ES009605 and R01 ES021369 and EPA grants RD 82670901 and RD 83451301.

CHS: The CHS was supported by the following NIH grants: 5K01ES017801, 1R01ES022216, 5P30ES007048, R01ES014447, P01ES009581, R826708-01 and RD831861-01.

EDEN: EDEN funding was provided by: Funds for Research in Respiratory Health, the French Ministry of Research: IFR program, INSERM Nutrition Research Program, French Ministry of Health: Perinatality Program, French National Institute for Population Health Surveillance (INVS), Paris–Sud University, French National Institute for Health Education (INPES), Nestlé, Mutuelle Générale de l'Education Nationale (MGEN), French speaking association for the study of diabetes and metabolism (Alfediam), grant # 2012/51290-6 Sao Paulo Research Foundation (FAPESP), EU funded MedAll project.

GALA II: The GALA II study was supported in part by grants from National Institutes of Health: the National Heart, Lung and Blood Institute (HL088133, HL078885, HL004464, HL104608, and HL117004); the National Institute of Environmental Health Sciences (ES015794 and ES24844); the National Institute on Minority Health and Health Disparities (MD006902); the National Institute of General Medical Sciences (GM007546); the American Asthma Foundation (E.G.B.); an RWJF Amos Medical Faculty Development Award (E.G.B.); the Sandler Foundation (E.G.B.); and the Flight Attendant Medical Research Institute (S.S.O.).

GECKO: The GECKO Drenthe birth cohort was funded by an unrestricted grant of Hutchison Whampoa Ld, Hong Kong. This methylation project in the GECKO Drenthe cohort was supported by the Biobanking and Biomolecular Research Infrastructure Netherlands [CP2011-19].

Generation R: The Generation R Study is made possible by financial support from the Erasmus Medical Center, Rotterdam, the Erasmus University Rotterdam and the Netherlands Organization for Health Research and Development. The EWAS data was funded by a grant to V.W.J. from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO) Netherlands Consortium for Healthy Aging (NCH; project nr. 050-060-810), and by funds from the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, Rotterdam, the Netherlands. V.W.J. received a grant from the Netherlands Organization for Health Research and Development (VIDI 016.136.361) and a Consolidator Grant from the European Research Council (ERC-2014-CoG-64916). L.D. received an additional grant from the Lung Foundation Netherlands (no 3.2.12.089; 2012). J.F.F. has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 633595 (Dynahealth).

INMA: INMA was funded by grants from Instituto de Salud Carlos III (Red INMA G03/176, CB06/02/0041), Spanish Ministry of Health (FIS-PI041436, FIS-PI081151), Generalitat de Catalunya-CIRIT 1999SGR 00241, Fundació La marató de TV3 (090430), EU Commission (261357-MeDALL: Mechanisms of the Development of ALLergy).

IOW: The IOW third generation cohort was funded by NIAID/NIH R01AI091905. This study was supported in part by NIAID/NIH R01AI091905 and R01AI121226.

MeDALL: The MeDALL study was supported by the European Union under the Health Cooperation Work Programme of the 7th Framework programme (grant agreement number 261357).

MoBa 1 and 2: This research was supported [in part] by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES-49019). The

Norwegian Mother and Child Cohort Study is supported by the Norwegian Ministry of Health and the Ministry of Education and Research, NIH/NIEHS (contract no. N01-ES-75558), NIH/NINDS (grant no.1 UO1 NS 047537-01) and the Norwegian Research Council/FUGE (grant no. 151918/S10).

MoBa3: MoBa3 data was funded by INCA/Plan Cancer, France. The work performed by the Epigenetics Group at the International Agency for Research on Cancer (IARC, Lyon, France) was supported by the grant from INCa/INSERM (France) to Z.H. and a Postdoctoral Fellowship (to A.G.) from IARC, partially supported by the EC FP7 Marie Curie Actions-People-Co-funding of regional, national and international programmes (COFUND) and the International Childhood Cancer Cohort Consortium (I4C).

NEST: The NEST study was funded by NIEHS grants R21ES014947 and R01ES016772 and NIDDK grant R01DK085173.

NFCS: This research was supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01 ES044005, ES049033, ES049032).

PIAMA: The PIAMA study was supported by The Netherlands Organization for Health Research and Development; The Netherlands Organization for Scientific Research; The Netherlands Asthma Fund; The Netherlands Ministry of Spatial Planning, Housing, and the Environment; and The Netherlands Ministry of Health, Welfare, and Sport.

Rotterdam gene expression: The Rotterdam Study is funded by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. Abbas Dehghan is supported by Netherlands Organisation for Scientific Research (NWO) grant (veni, 916.12.154) and the EUR Fellowship.

SEED: The SEED study was funded by the Centers for Disease Control and Prevention (grant nos. U10DD000180, U10DD000181, U10DD000182, U10DD000183, U10DD000184, U10DD000498) and the methylation assays were funded by Autism Speaks (grant no. 7659)

Project Viva: The Project Viva cohort is funded by NIH grants R01HL111108, R01NR013945, and R37 HD034568.

Supplemental References

1. Boyd, A. *et al.* Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children. *Int J Epidemiol* **42**, 111-27 (2013).
2. Fraser, A. *et al.* Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol* **42**, 97-110 (2013).
3. Touleimat, N. & Tost, J. Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. *Epigenomics* **4**, 325-41 (2012).
4. Wickman, M., Kull, I., Pershagen, G. & Nordvall, S.L. The BAMSE project: presentation of a prospective longitudinal birth cohort study. *Pediatr Allergy Immunol* **13 Suppl 15**, 11-13 (2002).
5. Chen, Y.A. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203-9 (2013).
6. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).
7. Eskenazi, B. *et al.* CHAMACOS, a longitudinal birth cohort study: lessons from the fields. *J Childrens Health* **1**, 3–27 (2003).
8. Eskenazi, B. *et al.* Association of in utero organophosphate pesticide exposure and fetal growth and length of gestation in an agricultural population. *Environmental Health Perspectives* **112**, 1116-1124 (2004).
9. Yousefi, P. *et al.* Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* **8**, 1141-52 (2013).
10. Teschendorff, A.E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189-96 (2013).
11. McConnell, R. *et al.* Traffic, susceptibility, and childhood asthma. *Environ Health Perspect* **114**, 766-72 (2006).
12. Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510-22 (2010).
13. Triche, T.J., Jr., Weisenberger, D.J., Van Den Berg, D., Laird, P.W. & Siegmund, K.D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* **41**, e90 (2013).
14. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-93 (2003).
15. Kumar, R. *et al.* Factors associated with degree of atopy in Latino children in a nationwide pediatric sample: the Genes-environments and Admixture in Latino Asthmatics (GALA II) study. *J Allergy Clin Immunol* **132**, 896-905 e1 (2013).
16. Nishimura, K.K. *et al.* Early-life air pollution and asthma risk in minority children. The GALA II and SAGE II studies. *Am J Respir Crit Care Med* **188**, 309-18 (2013).
17. Oh, S.S. *et al.* Effect of secondhand smoke on asthma control among black and Latino children. *J Allergy Clin Immunol* **129**, 1478-83 e7 (2012).
18. R Core Team (2013). A language and environment for statistical computing. Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/.
19. Aryee, M.J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363-9 (2014).
20. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E. & Storey, J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882-3 (2012).

21. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3**, 771-84 (2011).
22. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol* **13**, R44 (2012).
23. L'Abee, C. *et al.* Cohort Profile: the GECKO Drenthe study, overweight programming during early childhood. *Int J Epidemiol* **37**, 486-9 (2008).
24. Jaddoe, V.W. *et al.* The Generation R Study: design and cohort update 2012. *Eur J Epidemiol* **27**, 739-56 (2012).
25. Kruithof, C.J. *et al.* The Generation R Study: Biobank update 2015. *Eur J Epidemiol* **29**, 911-27 (2014).
26. Wang, D. *et al.* IMA: An R package for high-throughput analysis of Illumina's 450K Infinium methylation data. *Bioinformatics* (2012).
27. Guxens, M. *et al.* Cohort Profile: the INMA--INfancia y Medio Ambiente--(Environment and Childhood) Project. *Int J Epidemiol* **41**, 930-40 (2012).
28. Baiz, N. *et al.* Cord serum 25-hydroxyvitamin D and risk of early childhood transient wheezing and atopic dermatitis. *J Allergy Clin Immunol* **133**, 147-53 (2014).
29. Wijga, A.H. *et al.* Cohort profile: the prevention and incidence of asthma and mite allergy (PIAMA) birth cohort. *Int J Epidemiol* **43**, 527-35 (2014).
30. Gagnon-Bartsch, J.A. & Speed, T.P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13**, 539-52 (2012).
31. Zhang, B. *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. *Cell* **153**, 707-20 (2013).
32. Magnus, P. *et al.* Cohort profile: the Norwegian Mother and Child Cohort Study (MoBa). *Int J Epidemiol* **35**, 1146-50 (2006).
33. Ronningen, K.S. *et al.* The biobank of the Norwegian Mother and Child Cohort Study: a resource for the next 100 years. *Eur J Epidemiol* **21**, 619-25 (2006).
34. Haberg, S.E. *et al.* Maternal folate levels in pregnancy and asthma in children at age 3 years. *J Allergy Clin Immunol* **127**, 262-4, 264 e1 (2011).
35. Joubert, B.R. *et al.* 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ Health Perspect* **120**, 1425-31 (2012).
36. Midttun, O., Hustad, S. & Ueland, P.M. Quantitative profiling of biomarkers related to B-vitamin status, tryptophan metabolism and inflammation in human plasma by liquid chromatography/tandem mass spectrometry. *Rapid Commun Mass Spectrom* **23**, 1371-9 (2009).
37. Shaw, G.M. *et al.* Mid-pregnancy cotinine and risks of orofacial clefts and neural tube defects. *J Pediatr* **154**, 17-9 (2009).
38. Joubert, B.R. *et al.* Maternal smoking and DNA methylation in newborns: in utero effect or epigenetic inheritance? *Cancer Epidemiol Biomarkers Prev* **23**, 1007-17 (2014).
39. Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-95 (2011).
40. Johnson, W.E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-27 (2007).
41. Wilcox, A.J. *et al.* Folic acid supplements and risk of facial clefts: national population based case-control study. *BMJ* **334**, 464 (2007).
42. Markunas, C.A. *et al.* Identification of DNA Methylation Changes in Newborns Related to Maternal Smoking during Pregnancy. *Environ Health Perspect* (2014).
43. Davis, S., Du, P., Bilke, S., Triche, T., Jr. & Bootwalla, M. (2015). methylumi: Handle Illumina methylation data.

44. Hofman, A. *et al.* The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* **28**, 889-926 (2013).
45. Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692-702 (2011).
46. Westra, H.J. *et al.* MixupMapper: correcting sample mix-ups in genome-wide datasets increases power to detect small genetic effects. *Bioinformatics* **27**, 2104-11 (2011).
47. Schurmann, C. *et al.* Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One* **7**, e50938 (2012).
48. Schendel, D.E. *et al.* The Study to Explore Early Development (SEED): a multisite epidemiologic study of autism by the Centers for Autism and Developmental Disabilities Research and Epidemiology (CADDRE) network. *J Autism Dev Disord* **42**, 2121-40 (2012).
49. Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet* **3**, 1724-35 (2007).
50. Oken, E. *et al.* Cohort profile: project viva. *Int J Epidemiol* **44**, 37-48 (2015).