DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis

Bonnie R Joubert,[1,52] Janine F Felix,[2-4,52] Paul Yousefi,[5,52] Kelly M Bakulski,[6,52] Allan C Just,[7,52] Carrie Breton,[8,52] Sarah Reese,[1,52] Christina Markunas,[1,9,52] Rebecca C Richmond,[10,52] Cheng-Jian Xu,[11-13,52] Leanne Küpers,[14,52] Sam Oh,[15,52] Cathrine Hoyo,[16,52] Olena Gruzieva,[17,18,52] Cilla Söderhäll,[19,52] Lucas A Salas,[20-22,52] Nour Baïz,[23,52] Hongmei Zhang,[24,52] Johanna Lepeule,[25] Carlos Ruiz,[20-22] Symen Ligthart,[2] Tianyuan Wang,[1] Jack Taylor,[1] Liesbeth Duijts,[2,26,27] Gemma C Sharp,[10] Soesma A Jankipersadsing,[11,12] Roy M Nilsen,[28] Ahmad Vaez,[14] M Daniele Fallin,[29] Donglei Hu,[28] Augusto A. Litonjua,[30] Bernard F Fuemmeler,[31] Karen Huen,[5] Juha Kere,[19] Inger Kull,[18] Monica Cheng Munthe-Kaas,[32] Ulrike Gehring,[33] Mariona Bustamante,[17,20-22] Marie José Saurel-Coubizolles,[34] Bilal M Quraishi,[24] Jie Ren,[8] Jorg Tost[35] Juan R Gonzalez,[20-22] Marjolein J Peters,[36] Siri E Håberg,[37] Zongli Xu,[1] Joyce B van Meurs,[36] Tom R Gaunt,[10] Marjan Kerkhof,[14] Eva Corpeleijn,[14] Andrew P Feinberg,[6] Celeste Eng,[28] Andrea A Baccarelli,[7] Sara E Benjamin Neelon,[29] Asa Bradman,[5] Simon Kebede Merid,[18] Anna Bergström,[18] Zdenko Herceg,[38] Hector Hernandez-Vargas,[38] Bert Brunekreef,[33] Mariona Pinart,[20-22,39] Barbara Heude,[40,41] Susan Ewart,[42] Jin Yao,[8] Nathanaël Lemonnier,[43] Oscar H Franco,[2] Michael C Wu,[44] Albert Hofman,[2] Wendy McArdle,[10] Peter Van der Vlies,[11] Fahimeh Falahi,[14] Matthew W Gillman,[45] Lisa F Barcellos,[5] Ashish Kumar,[18] Magnus Wickman,[18] Stefano Guerra,[20] Marie-Aline Charles,[40] John Holloway,[46] Charles Auffray,[43] Henning W Tiemeier,[4] George Davey Smith,[10] Dirkje Postma,[13] Marie-France Hivert,[45] Brenda Eskenazi,[5] Martine Vrijheid,[20-22] Hasan Arshad,[47] Josep M Antó,[20-22,39] Abbas Dehghan,[2] Wilfried Karmaus,[24,53] Isabella Annesi-Maesano,[48,53] Jordi Sunyer,[20-22,39,53] Akram Ghantous,[38,53] Göran Pershagen,[18,53] Nina Holland,[5,53] Susan Murphy,[49,53] Dawn L DeMeo,[30,53] Esteban G Burchard,[15,50,53] Christine Ladd-Acosta,[29,53] Harold Snieder,[14,53] Wenche Nystad,[37,53] Gerard H Koppelman,[51,53] Caroline L Relton,[10,53] Vincent WV Jaddoe,[2-4,53] Allen Wilcox,[1,53] Erik Melén,[18,53] Stephanie J London*[1,53]

1. National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, North Carolina, 27709, USA.
2. Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
3. Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
4. The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
5. CERCH, School of Public Health, University of California Berkeley, Berkeley, CA, 94720-7360, USA.
6. Johns Hopkins University School of Medicine, Baltimore, MD, 21205, USA.
7. Department of Environmental Health, Harvard School of Public Health, Boston, MA, 02115, USA.
8. University of Southern California, Los Angeles, CA, 90032, USA.
9. Duke Molecular Physiology Institute, Duke University Medical Center, Durham, NC, 27710, USA.
10. MRC Integrative Epidemiology Unit, School of Social and Community Medicine, University of Bristol, Bristol, BS8 2BN, UK.
11. Department of Genetics, University Medical Center Groningen, University of Groningen, Gronginen, 9700 RB, the Netherlands.
12. Department of Pulmonology, University Medical Center Groningen, University of Groningen, Gronginen, 9700 RB, the Netherlands.
13. GRIAC Research Institute, Groningen, 9700 RB, the Netherlands.
14. Department of Epidemiology, University Medical Center Groningen, University of Groningen, Groningen, 9700 RB, the Netherlands.
15. Department of Medicine, University of California, San Francisco, San Francisco, CA, 94143-2911, USA.
16. Department of Biological Sciences and Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, 27695-7633, USA.
17. Center for Genomic Regulation (CRG), Barcelona, 08003, Spain.
18. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, SE-171 77, Sweden.
19. Department of Biosciences and Nutrition, Karolinska Institutet, Stockholm, SE-141 83, Sweden.
20. Centre for Research in Environmental Epidemiology (CREAL), Barcelona, 08003, Spain.
21. CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain.
22. Universitat Pompeu Fabra (UPF), Barcelona, 08003, Spain.
23. Department of Epidemiology of Allergic and Respiratory Department, Université Pierre et Marie Curie, 75654 Paris, France.

24. Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, 38152, USA.
25. Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, University of Grenoble Alpes, 38000 Grenoble, France.
26. Department of Pediatrics, Division of Neonatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
27. Department of Pediatrics, Division of Respiratory Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
28. Department of Global Public Health and Primary Care, University of Bergen, Bergen, 5018, Norway.
29. Johns Hopkins University Bloomberg School of Public Health, Baltimore, MD, 21205, USA.
30. Brigham and Women's Hospital, Channing Division of Network Medicine, Boston, MA, 02115, USA.
31. Department of Community and Family Medicine, Duke University School of Medicine, Durham, NC, 27710, USA.
32. Department of Pediatrics, Oslo University Hospital, Oslo, 424, Norway.
33. Institute for Risk Assessment Sciences, Utrecht University, Utrecht, 3508 TD, the Netherlands.
34. Epidemiological Research Unit on Perinatal Health and Women's and Children's Health, Institut National de la Santé et de la Recherche Médicale (INSERM), 75654 Paris, France.
35. Laboratory for Epigenetics and Environment, CEA Institut de Génomique, 91000 Evry, France.
36. Department of Internal Medicine, Erasmus MC, University Medical Center Rotterdam, Rotterdam, 3000 CA, the Netherlands.
37. Department of the Director for General Staff, Norwegian Institute of Public Health, Oslo, 0403, Norway.
38. Epigenetics Group, International Agency for Research on Cancer (IARC), 69008 Lyon, France.
39. Hospital del Mar Medical Research Institute (IMIM), Barcelona, 08003, Spain.
40. Early Origin of the Child's Health And Development (ORCHAD), Institut National de la Santé et de la Recherche Médicale (INSERM), 75654 Paris, France.
41. Epidemiology and Statistics Sorbonne Paris Cité Research Center (CRESS), Institut National de la Santé et de la Recherche Médicale (INSERM), 75654 Paris, France.
42. Department of Large Animal Clinical Sciences, Michigan State University, East Lansing, MI, 48824, USA.
43. European Institute for Systems Biology and Medicine, Université de Lyon, 69007 Lyon, France.
44. Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, USA.
45. Department of Population Medicine, Harvard Medical School, Boston, MA, 02215, USA.
46. Faculty of Medicine, Clinical & Experimental Sciences, University of Southampton, Southampton, SO16 6YD, UK.
47. Faculty of Medicine, Human Development & Health, University of Southampton, Southampton, SO16 6YD, UK.
48. Department of Epidemiology of Allergic and Respiratory Diseases, Institut National de la Santé et de la Recherche Médicale (INSERM), 75654 Paris, France.
49. Departments of Obstetrics and Gynecology and Pathology, Duke University School of Medicine, Durham, NC, 27710, USA.
50. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, San Francisco, CA, 94143-2911, USA.
51. Department of Pediatric Pulmonology and Pediatric Allergology, University Medical Center Groningen, University of Groningen, Groningen, 9700 RB, the Netherlands.

52. Equal contribution as first authors

53. Equal contribution as senior authors

* Correspondence to: Stephanie J. London NIEHS PO Box 12233 MD A3-05 Research Triangle Park NC 27709
london2@niehs.nih.gov

Abstract

Epigenetic modifications, including DNA methylation, represent a potential mechanism for environmental impacts on human disease. Maternal smoking in pregnancy remains an important public health problem that impacts child health in myriad ways with potential lifelong consequences. Mechanisms are largely unknown but epigenetics likely plays a role. We formed the Pregnancy And Childhood Epigenetics (PACE) consortium and meta-analyzed, across 13 cohorts (N=6,685), the association between maternal smoking in pregnancy and newborn blood DNA methylation at over 450,000 CpG sites using the Illumina 450K Beadchip. Over 6,000 CpGs were differentially methylated in relation to maternal smoking at genome-wide statistical significance (False Discovery Rate 5%), including 2,965 CpGs corresponding to 2,017 genes not previously related to smoking and methylation in either newborns or adults. Some genes are relevant to diseases that can be caused by maternal smoking (e.g. orofacial clefts and asthma) or adult smoking (e.g. certain cancers). A number of differentially methylated CpGs were associated with gene expression. We observed enrichment in pathways and networks critical in development. In older children (5 cohorts, N=3,187) 100% of CpGs gave at least nominal levels of significance, far more than expected by chance (p value < 2.2×10^{-16}). Results were robust to different normalization methods used across studies and cell type adjustment. In this large scale meta-analyses of methylation data, we identified numerous *loci* involved in response to maternal smoking in pregnancy with persistence into later childhood and provide insights into mechanisms underlying effects of this important exposure.

Introduction

Despite years of advisories regarding health risks to the developing fetus from maternal smoking, many pregnant women still smoke: 12.3% in the U.S.{Tong, 2013 #28} Maternal smoking during pregnancy is regarded as a cause of low birth weight, reduced pulmonary function (PLF [MIM: 608852]), orofacial clefts (OFC1 [MIM: 119530]), and sudden infant death syndrome (SIDS [MIM: 272120]) in exposed newborns.{US Department of Health and Human Services, 2014 #27} Other adverse birth outcomes{Moritsugu, 2007 #95} have been associated with maternal smoking in pregnancy along with common health problems in children including asthma (ASRT [MIM: 600807]), otitis media (OMS [MIM: 166760]), and neurobehavioral disorders.{US Department of Health and Human Services, 2014 #27}

The mechanisms for the adverse health effects of maternal smoking during pregnancy on offspring remain poorly understood.{US Department of Health and Human Services, 2014 #27} Recently, studies have examined the potential role of epigenetic modifications such as DNA methylation at specific CpG sites (CpGs). These include studies examining genome-wide DNA methylation in newborns in relation to maternal smoking in pregnancy using the Illumina Infinium HumanMethylation27 (27K) BeadChip{Breton, 2014 #102;Flom, 2011 #4;Suter, 2011 #8} or the newer platform with wider coverage, HumanMethylation450 (450K) BeadChip.{Joubert, 2012 #5;Markunas, 2014 #7;Richmond, 2015 #159;Kupers, 2015 #99} A number of differentially methylated *loci* have been identified in offspring in relation to maternal smoking in pregnancy in individual studies (references in Supplemental Note). One study examined the top CpGs with respect to timing of exposure and found that the signals reflect sustained, rather than short-term, exposure to maternal smoking during pregnancy{Joubert, 2014 #10} but this has not been evaluated genome-wide. A few studies suggest that some of these methylation signals persist into later childhood and adolescence but data are limited.{Lee, 2015 #91;Richmond, 2015 #159} Combining genome-wide data across studies using meta-analysis to generate large sample

sizes for the discovery of *loci* that would not have been identified from individual studies has been very successful in genetics, but this approach has rarely been used with methylation data.

To address the impact of maternal smoking during pregnancy on newborns with much greater power, we recruited 13 birth cohort studies with data on maternal smoking during pregnancy and DNA methylation in offspring from the 450K Beadchip into the Pregnancy and Child Epigenetics consortium (PACE). We meta-analyzed harmonized cohort-specific associations between maternal smoking during pregnancy and DNA methylation in the offspring. We examined both sustained maternal smoking and any smoking during pregnancy. We also examined persistence of DNA methylation patterns related to maternal smoking in newborns among older children, including adjustment for postnatal secondhand tobacco smoke exposure. For functional follow-up of findings we evaluated the associations between methylation status in the newly identified CpG sites and expression levels of nearby genes and performed pathway and functional network analyses. This study represents a large and comprehensive approach to evaluating the impact of maternal smoking during pregnancy on DNA methylation in offspring.

Material and Methods

Participating cohorts

A total of 13 PACE cohorts participated in the meta-analysis of maternal smoking during pregnancy and 450K DNA methylation in newborns. These studies, listed in alphabetical order, are the Avon Longitudinal Study of Parents and Children (ALSPAC), the Center for Health Assessment of Mothers and Children of Salinas (CHAMACOS), the Children's Health Study (CHS), the GECKO Drenthe cohort, the Generation R Study, Isle of Wight (IOW), Mechanisms of the Development of Allergy (MeDALL), three independent datasets from the Norwegian Mother and Child Cohort Study (MoBa1, MoBa2, and MoBa3), the Norway Facial Clefts Study

(NFCS), the Newborn Epigenetics Study (NEST), and Project Viva (VIVA). MeDALL represents a pooled analysis of four cohorts with coordinated methylation measurements: Infancia y Medio Ambiente (INMA), Etudes des Déterminants pré et postnatals précoces du développement et de la santé de l'ENfant (EDEN), Children's Allergy Environment Stockholm Epidemiology study (BAMSE), and Prevention and Incidence of Asthma and Mite Allergy (PIAMA). Two of the MeDALL cohorts contributed to the newborn meta-analysis (INMA and EDEN). There were five studies with data on older children: ALSPAC, Genes-environments and Admixture in Latino Americans (GALA II), the Study to Explore Early Development (SEED), MeDALL (INMA, EDEN, BAMSE, and PIAMA), and an independent methylation dataset from BAMSE subjects. The study methods for each cohort are described in detail in the supplemental material (see Supplemental Note).

Harmonization of maternal smoking variables

Cohorts assessed maternal smoking during pregnancy using questionnaires completed by the mothers. The MoBa study (MoBa 1 and MoBa 2) also used cotinine measurements from maternal blood samples taken during pregnancy as part of the definition of maternal smoking during pregnancy. More details on the cohort-specific smoking variables are in the Supplemental Note. In a previous publication from the MoBa1 study, significant associations between maternal smoking during pregnancy and DNA methylation in newborns were driven not by transient smoking that ended early in pregnancy but rather by sustained smoking during pregnancy.{Joubert, 2014 #10} Thus, each cohort ran separate models to evaluate both sustained smoking and any smoking during pregnancy. The sustained smoking during pregnancy (yes/no) variable was designed to capture women who smoked at least one cigarette per day through most of pregnancy. To cleanly contrast the effect of sustained smoking through pregnancy to never smoking during pregnancy, we excluded women who reported quitting smoking during pregnancy from the sustained smoking models. The any maternal smoking

during pregnancy (yes/no) variable was designed to capture any amount of smoking during pregnancy, at any time, even if a woman reported quitting. Because we did not exclude women who quit smoking during pregnancy from the any smoking during pregnancy models, the total sample sizes are slightly larger compared to the sustained smoking during pregnancy models. Genome-wide analyses use large sample statistics. We limited meta-analyses to cohorts with at least 15 subjects in both the exposed and unexposed groups. This excluded four cohorts (CHAMACOS, CHS, IOW, and VIVA) from the sustained smoking models. However these cohorts did participate in the any smoking during pregnancy meta-analysis.

Methylation measurements and quality control

Each cohort independently conducted laboratory measurements and quality control. The samples for each cohort underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, USA). Samples were processed with the Illumina Infinium HumanMethylation450 (450K) BeadChip (Illumina Inc., San Diego, USA) at Illumina or in cohort-specific laboratories.

Quality control of samples was performed by each cohort and included the exclusion of failed samples based on Illumina's detection p value, low sample DNA concentration, sample call rate, CpG specific percentage of missing values, bisulfite conversion efficiency, gender verification with multidimensional scaling plots, and other quality control metrics specific to cohorts. Cohorts could also use validated, published statistical methods for normalizing their methylation data on the untransformed methylation beta values (ranging from 0 to 1). Some cohorts also made independent probe exclusions. More details are provided in the Supplemental Note. For the meta-analysis, additional probe exclusions were made across all cohorts. Specifically, we excluded control probes (N=65), probes that mapped to the X (N=11,232) or Y (N=416) chromosomes, probes with an underlying SNP mapping to the last five nucleotides of the probe sequence (N=9,168) as previously described{Joubert, 2012 #5} and

CpGs with an implausible (zero) value for the standard error (N=67). This left a total of 464,628 CpGs included in the meta-analysis.

Cohort-specific statistical analyses

Each cohort ran independent statistical analyses according to a common pre-specified analysis plan. Robust linear regression was used in *R*{R Core Team, 2013 #11} to evaluate the association between maternal smoking during pregnancy and cord blood DNA methylation for each probe while accounting for potential heteroskedasticity and/or influential outliers. Each cohort ran the following covariate-adjusted statistical models: 1) the primary model, using sustained maternal smoking during pregnancy as the exposure and the normalized betas as the outcome, 2) sustained maternal smoking during pregnancy as the exposure and raw betas (not normalized) as the outcome, 3) any maternal smoking during pregnancy as the exposure and normalized betas as the outcome, 4) any maternal smoking during pregnancy as the exposure and raw betas as the outcome, and 5) sustained maternal smoking during pregnancy as the exposure and normalized betas as the outcome, with additional adjustment for cell type proportion. All models were adjusted for maternal age, maternal education (or a surrogate socioeconomic metric), parity, and technical covariates such as batch or plate. Some cohorts used *ComBat*{Leek, 2012 #52} to account for batch effects, and therefore did not include batch/plate as covariates in the models with normalized betas (see Supplemental Note). Additional correction for study design/sampling factors was done as needed in some cohorts. Because maternal smoking during pregnancy is not related to the child's sex, it cannot be a confounder and thus was not included in models. We did not adjust for principal components (PCs) because not all cohorts had genome wide genotype data and cohorts with genotype data had it only for a subset of subjects with methylation data. Furthermore, in one large cohort with PC data, models adjusted for PCs compared to models without PCs showed little variation in the results (correlation of betas =0.991; correlation of log(P-values)=0.996), despite a reduction

in sample size. The statistical models for cohorts with DNA methylation measured in older children were the same, with the additional adjustment for second-hand tobacco smoke exposure.

All cohorts independently estimated cell type proportion using the reference-based Houseman method{Houseman, 2012 #13} in the *minfi* package{Aryee, 2014 #55} using the Reinius et al. dataset for reference.{Reinius, 2012 #56} Cell type correction was applied by including the six estimated cell type proportions (CD8T, CD4T, NK cells, B cells, monocytes, granulocytes) as covariates in cohort-specific statistical models.

Meta-analysis

We performed inverse variance-weighted fixed-effects meta-analysis with *METAL*.{Willer, 2010 #16} Multiple testing was accounted for by controlling the false discovery rate (FDR) at 5%, implementing the method by Benjamini and Hochberg.{Benjamini, 1995 #17} This method was applied to all instances of FDR correction described in this paper unless otherwise specified. CpGs with an FDR-corrected p value less than 0.05 were considered statistically significant. CpGs that were statistically significant based on the more stringent Bonferroni correction (uncorrected p value $<1.08 \times 10^{-7}$ to account for 464,628 tests) were also noted.

To determine the robustness of our models and findings, we performed an additional analysis removing the cohorts of non-European ancestry (Supplemental Table S1). We compared the effect estimates, standard errors, and the distribution of the p values for the model to the estimates for our primary model to evaluate the consistency of our findings.

Examination in older children of CpGs associated with smoking in cord blood

The FDR significant CpGs identified in the primary model from the newborn meta-analyses were followed up using a lookup replication approach in the results from five older children cohorts, applying FDR correction to account for the number of CpGs tested.

Literature review to identify genes previously associated with smoking and methylation

We performed a systematic literature review to determine which CpGs represented findings not previously related to smoking exposure and methylation in the literature. A query of NCBI's PubMed database was performed using the search terms (("DNA Methylation"[Mesh] OR methylation) AND ("Smoking"[Mesh] OR smoking)) to be broad enough to capture all past studies reporting such results. CpGs with previously reported associations with smoking, both from prenatal exposure or in adults, were considered. This search yielded 789 results when performed on March 1, 2015. All results were then reviewed by title and abstract to determine whether they met inclusion criteria. First, results were limited to those performed in healthy human populations. That is, participants could not exclusively have been drawn from disease cases, such as cancer patients, and could not have been performed only in cell lines or animal studies. Case-control analyses that included healthy controls were accepted as meeting this criterion and no limitation was applied concerning the tissue used for DNA extraction. Second, studies were required to have performed DNA methylation analysis agnostically on a large scale as opposed to targeted interrogation of candidate CpG sites. This was operationalized by including only analyses that examined >1,000 sites simultaneously. The Illumina 450K, 27K, and GoldenGate arrays all met this criterion. Third, the exposure was restricted to tobacco cigarette smoking. Related exposures, such as to other forms of tobacco use or smoke exposure, were not included. Lastly, studies had to have reported their significant results publicly. Studies that failed to report p values or gene annotations were excluded.

Review of the existing literature on the effect of smoking on DNA methylation identified 25 publications meeting inclusion criteria. Of these, 16 studies reported results for adult smoking

exposure,{Besingi, 2014 #111;Breitling, 2011 #112;Dogan, 2014 #92;Elliott, 2014 #114;Flanagan, 2015 #115;Guida, 2015 #116;Philibert, 2012 #118;Philibert, 2013 #117;Shenker, 2013 #58;Siedlinski, 2012 #120;Sun, 2013 #121;Tsaprouni, 2014 #122;Wan, 2012 #124;Wan, 2014 #123;Zaghlool, 2015 #125;Zeilinger, 2013 #59} while nine provided results of association between maternal smoking during pregnancy on child DNA methylation.{Breton, 2014 #102;Chhabra, 2014 #103;Harlid, 2014 #93;Ivorra, 2015 #88;Joubert, 2012 #5;Lee, 2015 #91;Maccani, 2013 #108;Markunas, 2014 #7;Richmond, 2015 #159} CpG level results (p values and gene annotations) for sites showing significant association between smoking exposure and DNA methylation were extracted and compiled for comparison with the results from the meta-analysis. Results were considered significant if they met the multiple testing criteria implemented within the publication. For studies failing to implement any multiple testing correction, a naive Bonferroni threshold for the number of tests performed in the individual study was used. Genes previously associated with either adult smoking or maternal smoking in pregnancy (Supplemental File 3) and were excluded from our list of meta-analysis results.

CpG annotation

The official gene name was noted for each CpG using Illumina's genome coordinate.{, 2014 #19} We enhanced the annotation provided by Illumina based on the RefSeq database with additional UCSC and Ensembl databases, as well as annotation data in Bioconductor. All of the annotations use the human Feb. 2009 (GRCh37/hg19) assembly. We also used the package Snipper to annotate the nearest genes within 10 Mb of each CpG. We include this expanded Snipper gene annotation in our tables and discussion.

For selected genes, we used *CoMet*{Martin, 2015 #138} to graphically display additional information about CpGs including physical location, correlation, statistical significance, and functional annotation.

334

335 *Enrichment analysis*

336 We evaluated whether the CpGs significantly associated with smoking (based on the
337 FDR p value < 0.05) were enriched, relative to all CpGs analyzed, for several biologic
338 annotations provided in the Illumina annotation file. We assessed enrichment using the two-
339 sided doubling mid p value of the hypergeometric test.[24] We also evaluated enrichment in a
340 subset of CpGs mapping to imprinted differentially methylated regions (DMRs) described by
341 Court et al.{Court, 2014 #31}

342

343 *Pathway analyses*

344 We linked the CpGs significantly associated with smoking (based on the FDR p value <
345 0.05) to genes based only on the 450K BeadChip annotation file.{Triche, 2014 #20} Probes
346 lacking an annotated *Entrez Gene* ID were filtered (n=1,971), as were duplicate gene entries
347 (n=1,473). A total of 2,629 unique gene identifiers were used in gene ontology enrichment
348 analysis with three different procedures as described below.

349 This resulted in 2,235 genes that mapped to gene ontologies of biological processes and
350 we tested for gene enrichment over the background array (16,119 unique annotated Entrez
351 Gene IDs) using Fisher's exact tests with a minimum of five genes per node using topGO in
352 *R*.{Alexa, 2010 #21} In addition, we used the DAVID bioinformatics resource{Huang da, 2009
353 #22} to test for enrichment in gene ontology biological processes with a threshold of five, using
354 the Benjamini-Hochberg procedure to control for false discoveries. Finally, we used QIAGEN's
355 Ingenuity Pathway Analysis to identify relevant signaling and functional pathways (IPA,
356 Redwood City, CA, www.qiagen.com/ingenuity).

357

358 *Functional network analysis*

To construct a functional association network, it was desirable to reduce the list of tested CpGs, so we prioritized the FDR significant CpGs from the primary model in a stepwise manner. First, we only included those CpGs that were FDR significant in both the primary and cell type adjusted models (FDR *P* values <0.05). Next, we sorted these CpGs based on their effect size (beta coefficient), and selected the top quartile (N=980). The genes mapping to these prioritized CpGs were then used as input for the construction of a functional interaction network. We used the GeneMANIA algorithm as well as its functional association data including genetic interaction, physical interactions, co-expression, shared protein domains, and co-localization networks.{Warde-Farley, 2010 #85} We selected the "all available networks" option with a 500-gene output (accessed March 11, 2015). Functional enrichment analysis was then performed on all genes from the constructed interaction network against Gene Ontology (GO) terms to detect significantly enriched GO terms.{Vaez, 2015 #86} FDR correction was applied to this analysis based on the q-value using a threshold of q<0.01.

Methylation transcription analysis

To further explore the associations between methylation and gene expression, we performed methylation-expression analyses, evaluating the association between methylation status of CpGs and differences in quantitative levels of gene expression. All identified CpGs that reached FDR-corrected significance and that we identified as not previously reported in the literature were tested for association with expression levels of genes within a region of 250 kb upstream or downstream of the CpG{Zhang, 2014 #160} (total region 500kb), to evaluate whether the CpG-methylation status influenced transcript levels of genes. We had two data sets available for this analysis. One dataset included mRNA gene expression (Illumina HumanHT12v4) and 450K methylation data, both from whole blood samples from 730 adults over 45 years of age in the Rotterdam Study, a population-based prospective cohort study in Rotterdam, the Netherlands.{Hofman, 2013 #127} This gene expression dataset is available at

GEO (Gene Expression Omnibus) public repository under the accession GSE33828. The second dataset included mRNA gene expression (Affymetrix HTA 2.0) and 450K methylation data on whole blood samples from 107 children aged 4 years from the INMA study in Spain. Study population details for the Rotterdam Study and INMA are in the supplementary note. In the Rotterdam dataset, 2,636 of the 2,965 CpGs examined mapped to a transcript within the 500kb window. We created residuals for mRNA expression after regressing out the Houseman estimated white blood cell proportions, the erythrocytes and platelets cell counts, fasting state, RNA quality score, plate number, age, and sex on the mRNA expression levels using a linear mixed model. We then created residuals for DNA methylation regressing out the Houseman estimated white blood cell proportions, age, sex, batch effects on the dasen normalized beta-values of the CpG sites using a linear mixed model. We used a linear regression model to evaluate the association between the residuals of the mRNA expression levels and the residuals of the dasen normalized beta-values of the CpG sites.

The INMA gene expression data was normalized using Expression Console Software from Affymetrix and probes were clustered to the transcript level. Only transcripts within the 500 kb window of selected CpG sites were considered in the analysis (N=45,076 transcripts). To control for technical variation in the DNA methylation dataset, a principal component (PC) analysis of 600 negative control probes using 10,000 permutations was performed, and the residuals of a linear regression model using the first 5 PCs were estimated. The effect of sex and Houseman cell proportions estimates were adjusted for in a second stage linear regression model. Two models were applied to control for technical and unwanted biological variation when estimating gene expression residuals. In the first one, sex and Houseman estimates were regressed out. In the second one, fourteen surrogate variables estimated using the sva R package were adjusted for in a second model including sex and Houseman estimates. A linear regression model of residuals of gene expression versus residuals of methylation was

410 performed. Multiple testing for both Rotterdam and INMA gene expression analyses was
411 controlled using Benjamini-Hochberg FDR correction.
412
413 *Examination of polymorphic and cross-reactive probes*
414 The list of FDR significant CpGs was matched to the list of polymorphic and cross-
415 reactive CpGs provided by Chen et al.{Chen, 2013 #18} to identify potential problematic probes.
416 We additionally performed the dip test{Hartigan, 1985 #161} for unimodality for each CpG to test
417 for non-unimodal distributions in the MoBa1 cohort (N=1,068). Also using the MoBa1 cohort, we
418 visually inspected density plots for each of the probes that matched to the list of polymorphic
419 probes from Chen et al. to assess departures from unimodality, including from small numbers of
420 outlier values.
421
422 **Results**
423 *Study characteristics*
424 A total of 13 cohorts participated in the meta-analysis of maternal smoking during
425 pregnancy and 450K DNA methylation in newborns. Among these 6,685 newborns, 897 (13%)
426 were exposed to sustained maternal smoking during pregnancy and 1,646 (25%) were exposed
427 to any maternal smoking during pregnancy. We also included five cohorts of older children
428 (N=3,187, average age = 6.8 years); 266 children (8%) were exposed to sustained smoking
429 during pregnancy and 404 (13%) were exposed to any maternal smoking during pregnancy. The
430 cohort-specific summary statistics for maternal smoking are presented in Table 1 and covariates
431 in Supplemental Table S1. The majority of participants were of European ancestry
432 (Supplemental Table S1).
433
434 *Meta-analysis*

Our primary model evaluated the association between sustained maternal smoking during pregnancy and differential DNA methylation in newborns using normalized methylation betas as the outcome, adjusting for covariates (Figure 1). The cohort-specific lambdas and number of CpGs included in each model are listed in Supplemental Table S2. Among the 6,073 CpGs with FDR significance (Supplemental Table S3), 568 also met the strict Bonferroni threshold for statistical significance (p value<1.08×10^{-7}, correcting for 464,628 independent tests). Results were quite robust to cell type adjustment (Supplemental Table S3): all 568 Bonferroni-significant CpGs from the primary model remained FDR-significant in the cell type adjusted model and 78% were Bonferroni significant in both models. The $\log_{10}$(p values) for the primary model and cell-type adjusted models were highly correlated (correlation coefficient = 0.92 across all CpGs, 0.98 for the CpGs FDR-significant in the primary model, Supplemental Figure 1). Given the general similarity of the results before and after cell type adjustment and the fact that the available reference panel is from a small number (N=6) of adult men,{Reinius, 2012 #56} we regard the covariate adjusted model as the primary model. The results for other models (the cell type adjusted model, the any smoking during pregnancy model, and the sustained smoking during pregnancy associated with older children methylation model) and the mean methylation values in newborns and older children are shown for all 6,073 CpGs in Supplemental Table S3.

Among the 6,073 FDR-significant CpGs, smoking during pregnancy was associated approximately equally with increased methylation (52%) and decreased methylation (48%) (Figure 2). Out of the 3,932 CpGs that were also FDR-significant after cell type adjustment, there were 967 CpGs in or within 10 Mb of the 1,185 genes we identified in our systematic literature review (see methods, Supplemental Note and Supplemental Table S4) as previously reported to be differentially methylated in relation to smoking. This left 2,965 CpGs (corresponding to 2,017 annotated mapped or nearest genes) that had not previously been reported (Supplemental Table S5; genes highlighted in discussion shown in Table 2). For

comprehensive comparison with the previous literature, we also present our results for all CpGs that were either not FDR-significant after cell type adjustment and/or that annotated to genes already described in the literature as related to smoking and methylation (N=3,108 CpGs, Supplemental Table S6). Our top finding among the 6,073 FDR-significant CpGs was for *AHRR* (MIM: 606517) cg05575921 (p value = 1.64×10^{-193}) which is the top most statistically significant CpG in many other studies evaluating either personal smoking or maternal smoking during pregnancy.

We found our results to be robust to different analytic approaches. We present results from models using normalized betas as the outcome. When using raw betas as the outcome we observed little difference in the results (Spearman's correlation coefficient = 0.96 for regression coefficients; 0.98 for $\log_{10}$(p values) for our significant findings). Furthermore, excluding the one cohort with newborns of non-European ancestry (NEST) from the sustained maternal smoking model provided similar results (Spearman's correlation coefficient = 0.99 for regression coefficients; 0.89 for $\log_{10}$(p values).

Examination of potentially polymorphic and cross-reactive probes

A total of 742 of the 6,073 FDR significant CpGs overlapped with the list of 70,889 potentially polymorphic probes provided by a supplemental table of Chen *et al.*{Chen, 2013 #18} Only 137 of the 6,073 FDR-significant CpGs overlapped with the list of 29,233 cross-reactive probes annotated by Chen et al. Many of the probes flagged by Chen are associated with very low frequency SNPs and thus are likely to have minimal impact on results in most datasets. In visual inspection of the density plots of all 742 such probes, we flagged 19 CpGs as having a possible deviation from unimodality (listed in Supplemental Table S7). However, results from the dip test{Hartigan, 1985 #161} applied to all 6,073 FDR significant CpGs identified only 4 CpGs as statistically significantly deviated from unimodality (FDR adjusted p<0.05; cg11459648, cg17847044, cg15028160, and cg25849281).

Persistence of DNA methylation related to maternal smoking during pregnancy in older children

Because of the smaller sample size and smaller proportion of children exposed to maternal smoking in the older children models, we had less statistical power compared to the newborn models. When we compared the coefficients for newborns and older children for all 6,073 CpGs, there were 4,403 (73%) with a consistent direction of effect and all 6,073 (100%) with nominal P values < 0.05 for the older children models which is higher than the 5% expected by chance alone (Kolmogorov p value < 2.2×10^{-16}). Among these, 3,722 CpGs (61%) had a weaker effect size (attenuation) in the older children compared to newborns, but the attenuation overall was very small in magnitude and not significant (mean attenuation=-0.00039, SD=0.0059). Among the 148 CpGs that met FDR significance at look-up replication level in the older children (Supplemental Table S8), 100% were consistent in the direction of effect compared to newborns and there was attenuation for 32%, again small in magnitude and not significant (mean attenuation=-0.00008, SD=0.018).

Enrichment analysis

For our 6,073 FDR significant CpGs, we observed enrichment for localization to CpG island shores (35% versus 23% overall as compared to all CpGs on the array, p value=2.8×10^{-100}), enhancers (29% versus 22% overall, p value=5.7×10^{-45}), and DNAse hypersensitivity sites (14% versus 12% overall, p value=2.8×10^{-7}). Conversely, we found relative depletion in CpG islands (18% versus 31% overall, p value=9.1×10^{-116}), FANTOM promoters (2.5% versus 6.7% overall, p value=2.1×10^{-49}), and promoter associated regions (13% versus 19% overall, p value=2.3×10^{-33}). There was no statistically significant enrichment or depletion of sites mapping to imprinted DMRs (0.082% versus 0.16% overall, p value=0.107).

Pathway analysis

Our pathway analyses indicated that the FDR significant CpG sites corresponded to genes enriched for several categories of biological processes including anatomical development, phosphate-containing compound metabolism, nervous system development, and cell communication processes (Supplemental Figure 2). Based on DAVID, eight biological processes were enriched including GTPase signal transduction, neuronal differentiation, and protein kinase activity (Supplemental Figure 3). The top statistically significantly enriched diseases and biofunctions identified through Ingenuity software included tumor adhesion, neuron development, and connective tissue differentiation (Supplemental Figure 4).

Functional network analysis

Functional network analysis revealed 447 significantly enriched GO terms after applying FDR correction (q-value <0.01 for this analysis, Supplemental Table S9). The majority of the enriched terms, and particularly the most statistically significant ones, pointed towards biological processes related to cell/tissue/organ development, proliferation, morphogenesis, differentiation, growth, and other biologically relevant processes. There were also several enriched processes related to embryonic morphogenesis/development.

Methylation transcription analysis

To assess transcriptional effects related to methylation differences, we investigated whether methylation status correlated with gene expression levels for our 2,965 CpGs associated with sustained maternal smoking in newborns that we identified through literature review as not having been previously reported. In the Rotterdam Study dataset of adults, out of the 2,636 (of the 2,965) CpGs that we were able to match to a gene transcript (+/- 250 kb), 254 unique CpGs (343 total CpG-gene transcript associations) were significantly associated with expression of a nearby gene in whole blood from adults (FDR p value < 0.05, Supplemental Table S10). We observed strong associations for several CpGs annotated to the same gene

and corresponding gene expression levels, most strikingly for *IL32* (MIM: 606001) with four CpGs associated with *IL32* expression, and *HOXB2* (MIM: 142967) with several CpGs related to *HOXB2* expression (lowest p value 2.38 x 10^{-72}, Supplemental Table S10). In the much smaller study of children at age four from INMA (N=107), 35 CpGs were associated with gene expression (FDR p value < 0.05). The following six genes had CpGs with methylation that was statistically significantly related to gene expression in both the Rotterdam Study adults and INMA children: *ENOSF1* (MIM: 607427)*, HOXB2* (MIM: 142967*, IL32* (MIM: 606001*, NLRP2* (MIM: 609364)*, PASK* (MIM: 607505)*,* and *TDRD9*. In both the adult and child datasets, for the majority of CpGs statistically significantly related with expression the direction was inverse (higher methylation, lower expression).

DNA methylation related to any maternal smoking during pregnancy in newborns

In addition to the sustained smoking model we meta-analyzed the effect of any maternal smoking during pregnancy on newborn methylation. As expected, based on previous literature,[12] we found that despite the much larger number of women with any smoking during pregnancy, there were fewer statistically significant findings for this less specific exposure (4,653 FDR-significant CpGs, Supplemental Table S3).

Discussion

We combined data across many studies in a large scale epigenome-wide meta-analysis to evaluate the association between maternal smoking during pregnancy and DNA methylation in offspring. We established the Pregnancy and Childhood Epigenetics (PACE) consortium to study this association and had 13 birth cohort studies from the US and Europe that measured CpG-specific DNA methylation across the epigenome in newborns with the same reproducible platform. Combining these studies resulted in the discovery of 6,073 statistically significant CpGs; 3,932 remained statistically significant after adjustment for cell type proportion. Our

565 results are remarkably robust to different modeling techniques. Findings were very similar when
566 using either the raw methylation betas or the normalized betas as the outcome. This is despite
567 the variety of data processing methods used across the cohorts for normalization and
568 corrections for technical variables such as batch (described in the Supplemental Note). This
569 consistency is reassuring given the range of published methods available for researchers to
570 apply to 450K DNA methylation data for the quality control, normalization, and adjustment for
571 technical variation. Furthermore, our main findings persisted after cell type adjustment
572 (Supplemental Table S3, Supplemental Figure 1).

573 As predicted, based on earlier evaluation of top findings for maternal smoking in the
574 MoBa cohort,[12] we had fewer statistically significant findings for any smoking during pregnancy
575 than for sustained smoking during pregnancy (Supplemental Figure 5). Nonetheless, with the
576 large sample size of this meta-analysis, we still observed many statistically significant CpGs
577 after FDR correction in the any smoking models and the direction of effect and p values were
578 similar to those from the sustained smoking models (Supplemental Table S3, Supplemental
579 Figure 6). However, the stronger signal for sustained smoking suggests that this might be the
580 more powerful variable for studying epigenetic effects and possible health outcomes from this
581 exposure in offspring.

582 Our observation of a large number of genome-wide significant CpGs related to maternal
583 smoking is not surprising given reports of multiple genome-wide significant *loci* identified in
584 single studies, all with smaller sample sizes.{Ivorra, 2015 #88;Joubert, 2012 #5;Markunas, 2014
585 #7;Dogan, 2014 #92;Harlid, 2014 #93;Kupers, 2015 #99} Reassuringly, among our myriad
586 findings, the top hit in all newborn models was *AHRR* cg05575921 (p value<1.64×10^{-193}) which
587 has been observed as differentially methylated in relation to smoking in many studies of adults
588 and children. {Joubert, 2012 #5;Markunas, 2014 #7;Monick, 2012 #57;Shenker, 2013
589 #58;Zeilinger, 2013 #59;Kupers, 2015 #99}

590 Our enrichment testing of the genome-wide results is in line with previous findings
591 showing that island shores, enhancers, and DNase I hypersensitive sites are more dynamic
592 (susceptible to methylation changes) than promoter regions{Ziller, 2013 #147} and imprinted
593 loci.{Ivanova, 2012 #148} These regions may be more resistant to changes in DNA methylation
594 in response to *in utero* exposure.{Ivanova, 2012 #148} Thus, it is not surprising that
595 associations between maternal smoking and newborn methylation may be more likely to be
596 found in island shore and enhancer regions as opposed to promoters or CpG islands.
597 To assess the underlying biology involved in the associated genomic regions, we applied
598 pathway and functional analyses as well as tests of enrichment. These results implicated
599 numerous neurological pathways, pathways involved in embryogenesis, and various
600 developmental pathways. These observations may provide insight into the etiology of childhood
601 health outcomes related to maternal smoking during pregnancy.
602 We focus discussion on some specific genes among the associations that based on our
603 systematic review had not been previously reported (2,965 CpGs annotating to 2,017 mapped
604 or nearest genes). For 27 of these genes, mutations or SNPs have been implicated in
605 susceptibility to orofacial clefts (as identified in the *Snipper* database review described in the
606 Supplemental Note). This includes the following genes (each representing 1 FDR significant
607 CpG unless otherwise specified): *BHMT2* (MIM: 605932, 6 CpGs), *GRHL3* (MIM: 608317),
608 *THADA* (MIM: 611800), *GAD67* (MIM: 605363), *TP63* (MIM: 603273,2 CpGs), *MSX1* (MIM:
609 142983), *WDR1* (MIM: 604734), *SPP1* (MIM: 166490), *BMP6* (MIM: 112266), *TFAP2A* (MIM:
610 107580), *COL11A2* (MIM: 120290,3 CpGs), *PDGFRA* (MIM: 173490), *MN1* (MIM: 156100),
611 *MSX2* (MIM: 123101, 4 CpGs), *PVT1* (MIM: 165140), *ZIC2* (MIM: 603073), *HOXA2* (MIM:
612 604685, 10 CpGs), *WNT3* (MIM: 165330), *RUNX2* (MIM: 600211, 2 CpGs), *TERT* (MIM:
613 187270), *SPATA13* (MIM: 613324, 2 CpGs), *VAX1* (MIM: 604294), *TIMP2* (MIM: 188825), *NOG*
614 (MIM: 602991), *BEST3* (MIM: 607337), *MYH9* (MIM: 160775), and *BMP4* (MIM: 112262, 6
615 CpGs) (results in Supplemental Table S5) . Although this does not imply that the smoking are

on the causal pathway, we note that the Surgeon General's Report summarizes the evidence as sufficient to infer a causal relationship between maternal smoking during pregnancy and these birth defects.{US Department of Health and Human Services, 2014 #27} Many of these genes also have varied biological effects relevant to other aspects of development.

Among this group of genes previously related to orofacial clefts, bone morphogenetic protein 4 (*BMP4*) is especially interesting. Maternal smoking might interact with SNPs in this gene in relation to oral clefts.{Chen, 2014 #97} We identified six CpGs in *BMP4* at genome-wide significance in newborns; two remained statistically significant in the older children. In addition to orofacial clefts, SNPs in *BMP4* are related to tooth development and eruption, as well as to colorectal cancer in GWAS.{Burdett, #94} BMPs, including BMP4,[45] also play an important role in lung development: reduced lung function among infants is an established consequence of maternal smoking during pregnancy.{US Department of Health and Human Services, 2014 #27} A plot showing greater detail on the CpGs in or near *BMP4* is provided in Figure 3.

We observed six CpGs significantly related to maternal smoking in betaine-homocysteine methyltransferase (*BHMT2*; Supplemental Figure 7). Genetic variants in this gene have been associated with orofacial clefts in candidate gene studies{Mostowska, 2010 #130} and with selenium levels in GWAS.{Welter, 2014 #131;Hindorff, #132} Of note, in experimental studies, selenium has been shown to protect against orofacial clefts induced by exposure to teratogens.{Ozolins, 1996 #133} In the Cancer Genome Atlas, methylation of *BHMT2* in lung adenocarcinoma (lung cancer [MIM: 211980]) was strongly correlated (3rd rank genome-wide) with smoking history.{, #134}

The gene *PRDM8* (PR domain containing 8, MIM: 616639) has the largest number of CpGs (18 of 61 based on Illumina annotation) significantly associated with maternal smoking during pregnancy. Maternal smoking during pregnancy was associated with decreased methylation throughout the gene. *PRDM8* is one of several *PRDMs* belonging to the SET domain family of histone methyltransferases.{Fog, 2012 #72} *PRDM* genes either act as direct

642 histone methyltransferases or recruit a suite of histone-modifying enzymes to target
643 promoters.{Hohenauer, 2012 #73} *PRDM8* specifically methylates H3K9 of histones to repress
644 transcriptional activity.{Eom, 2009 #71} *PRDM8* expression is tightly regulated in a spatio-
645 temporal manner during neural development;{Komai, 2009 #76} it regulates morphological
646 transition in neocortical development{Inoue, 2014 #74} and forms part of a repressor complex
647 that directs, through regulation of Cadherin-11, neural circuit assembly.{Ross, 2012 #80} Thus
648 *PRDM8* appears to play an important role in neurologic development.
649 *DLGAP2* (discs, large (Drosophila) homolog-associated protein 2, MIM: 605438) is
650 another gene with a large number of significant CpGs (14 of 192 tested CpGs) associated with
651 maternal smoking in our study. *DLGAP2* belongs to a gene family that encodes SAP90/PSD95-
652 associated proteins (SAPAPs), and SAPAP2 is known to be involved in the molecular
653 organization of synapses and in neuronal cell signaling.{Takeuchi, 1997 #70} *DLGAP2* was first
654 identified in studies of progressive epilepsy with mental retardation (EPMR [MIM:
655 610003]){Ranta, 2000 #69}, and has been associated with other central nervous system
656 disorders such as schizophrenia (SCZD [MIM: 181500]){Li, 2014 #66} and autism spectrum
657 disorders (ASD [MIM: 209850]).{Pinto, 2010 #68;Chien, 2013 #62} Differential methylation at
658 this locus in a rat model appears to play a role in the development of post-traumatic stress
659 disorder.{Chertkow-Deutsher, 2010 #61}
660 The neuropilin-2 (*NRP2,* MIM: 602070) gene had 3 CpGs located in close proximity
661 (among 48 tested) that were statistically significantly associated with maternal smoking during
662 pregnancy. *NRP2* is one of two transmembrane receptors for axonal guidance cues of the class
663 3 semaphorin (SEMA) family and is expressed in sympathetic neural crest cells and their
664 progeny.{Maden, 2012 #77} It may also be required *in vivo* for sorting migrating cortical and
665 striatal interneurons to their correct destination.{Wu, 2007 #82} *NRP2* also functions as a
666 receptor for some forms of vascular endothelial growth factor, thereby playing a crucial role in
667 angiogenesis and lymphangiogenesis.{Neufeld, 2002 #79} Polymorphisms in *NRP2* have been

associated with several diseases including autism{Wu, 2007 #82} and multiple cancers.{Nasarre, 2013 #78;Jubb, 2012 #75;Yasuoka, 2011 #83;Samuel, 2011 #81;Yasuoka, 2009 #84}

Hypermethylation of *ESR1* (estrogen receptor 1, MIM:133430; a key nuclear transcription factor) on chromosome 6q25.1 is well-studied in relation to presence and prognosis of various malignancies such as breast cancer and hepatocellular carcinoma{Gaudet, 2009 #65;Dai, 2014 #63} as well as asthma.{Dijkstra, 2006 #152;Koppelman, 2011 #153} We found an inverse association between maternal smoking and methylation levels for seven out of the eight FDR significant *ESR1* CpG sites. *ESR1* hypomethylation has been reported in relation to induced microRNA expression (synthetic miR-29b oligonucleotides) in acute myeloid leukemia cells.{Garzon, 2009 #64;Garzon, 2009 #64;Garzon, 2009 #64;Garzon, 2009 #64}

To evaluate possible functional gene expression effects of methylation at the CpGs that we found to be significantly related to maternal smoking, we analyzed data from two studies – one of adults and another of children at age four years. Although on first pass, one might expect a higher proportion of the CpGs related to maternal smoking to also be related to gene expression, there are several factors that decrease the likelihood of seeing significant associations. Most importantly, the sample size for discovery of the methylation association with smoking was much larger than the datasets available to correlate gene expression and methylation (about 10 fold smaller for the adult gene expression dataset and about 60 fold smaller for the childhood dataset). In addition, gene expression in blood may be more transient than methylation, decreasing the ability to find significant associations with a single gene expression measurement. Furthermore, constitutive gene expression is measured in this setting whereas many genes are inducible and methylation might contribute to this process. Lastly, some *in utero*-induced changes to methylation may have affected transcription during fetal development but not in postnatal life, and may have transcription-independent functional mechanisms. Nonetheless, we observed significant associations between methylation and gene

694 expression at six genes in both the adults and the children. The majority of CpGs significantly
695 associated with expression were in the commonly expected direction of methylation related to
696 gene silencing. Notably, CpGs in *IL32* (Supplemental Figure 8), a proinflammatory cytokine
697 involved in several diseases such as asthma{Meyer, 2012 #151} and cancer,{Joosten, 2013
698 #100}, *HOXB2*, a transcription factor involved in development{Tumpel, 2002 #157} and several
699 cancer forms{Boimel, 2011 #158} and *PASK* (PAS domain containing serine/threonine kinase),
700 involved in glucose homeostasis,{DeMille, 2013 #129} were significantly associated with
701 expression in both datasets.

702 We analyzed the associations of CpGs with expression levels of genes within a region of
703 250 kb up- or downstream of the CpG. Consensus on the optimal physical distance for these
704 analyses is lacking. However, in a recent study, associations between CpGs and SNPs were
705 the strongest when within close proximity (500 kb) of the CpG-site.{Zhang, 2014 #160} Despite
706 the limitations with the expression datasets included in our study, we believe that the
707 transcriptomics data provide functional support for our maternal smoking findings.

708 In older children, all of the CpGs significantly associated with maternal smoking in
709 newborns gave at least nominal levels of significance (p value < 0.05). This skew of the
710 distribution of *P* values toward small values was much more than expected by chance
711 (Kolmogorov p value < 2.2×10^{-16}) demonstrating a very high level of replication and persistence
712 of findings at birth into later childhood. This is consistent with and substantially extends a few
713 previous reports.{Lee, 2015 #91;Richmond, 2015 #159} We had only very limited data with
714 repeat measures in the same individuals so we could not meta-analyze change in methylation
715 over time.

716 This inaugural paper from the PACE consortium, represents a major effort to combine
717 data from many studies in a large scale meta-analyses of epigenome-wide association studies
718 of maternal smoking in relation to methylation in newborns. We report at least an order of
719 magnitude more genes differentially methylated in response to maternal smoking than have

been identified in any previous study. This suggests that meta-analysis in epigenome-wide association studies produces similar success to that of genome-wide association SNP studies in the identification of biologically meaningful *loci*. The similarity in the results obtained using the raw betas compared to normalized betas generated using various methods indicates that cohort-specific processing methods do not interfere with the ability to perform meta-analysis.

We identified nearly 3,000 CpGs corresponding to genes differentially methylated in offspring in relation to whether mothers smoked during pregnancy. Some of these genes have been implicated in genetic studies of orofacial clefts or asthma, both conditions related to maternal smoking in pregnancy and others in the pathogenesis of cancers that are associated with adult smoking, including lung, colorectal (CRC [MIM: 114500]) and liver (HCC [MIM: 114550])[2]. We also find substantial persistence into later childhood of effects of maternal smoking identified in newborns. Our findings may implicate epigenetic mechanisms in the etiology of these exposure-disease relationships. This large scale study also provides confirmation of previously reported loci, many of which have not been previously replicated. Pathway analysis highlights the involvement of identified genes in various developmental pathways, and functional effects at the transcriptomics level were observed for many of the identified CpG sites. These findings may provide new insights into the mechanisms involved in the detrimental health outcomes that arise from this important *in utero* exposure.

Supplemental Data Description

Supplemental Data include 8 figures, 10 tables (6 in Excel format), and the supplemental note (supplemental materials and methods, acknowledgements, and funding information).

Acknowledgements

For all studies, information on funding and acknowledgements can be found in the Supplemental Data.

Web Resources

The URLS for data presented herein are as follows:

Snipper, http://csg.sph.umich.edu/boehnke/snipper

The R Project for Statistical Computing, R v.3.0.2, http://www.r-project.org/

Infinium HumanMethylation450K v1.2 Product Files, http://support.illumina.com/downloads/infinium_humanmethylation450_product_files.html

A Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/gwastudies

References

Figure Legends

Figure 1. Meta-analysis of the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood. A total of 6,073 CpGs were considered statistically significant using FDR correction (solid horizontal line); 568 Bonferroni significant (dashed horizontal line)).

Figure 2. Volcano plot indicating the direction of effects for the meta-analysis of the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood.

Figure 3. Meta-analysis results for the association between sustained maternal smoking during pregnancy and DNA methylation in newborn cord blood: CpGs in or near *BMP4*. Top pane: – log10(P values) from the meta-analysis model, CpGs indicated by dots, color coded based on

771 pairwise correlation with neighboring CpGs. Middle panel: Annotation tracks for the plotted
772 genomic region. Bottom panel: Pairwise correlation matrix across the displayed CpGs.

Tables

Table 1. Smoking variable frequencies for the cohorts participating in meta-analyses: Newborns and older children

Study	Study population	N[a]	Exposed to sustained maternal smoking during pregnancy N (%)	Exposed to any maternal smoking during pregnancy N (%)
ALSPAC	newborns	860	87 (10.1)	120 (14.0)
CHAMACOS	newborns	378	7 (1.9)[b]	24 (6.3)
CHS	newborns	85	NA[b]	22 (25.9)
GECKO	newborns	255	70 (27.5)	129 (50.6)
Generation R	newborns	883	129 (14.6)	220 (24.9)
IOW	newborns	90	9 (10.0)[b]	23 (25.6)
MeDALL	newborns	362	43 (11.9)	63 (17.5)
MOBA1	newborns	1,063	156 (14.7)	312 (29.4)
MOBA2	newborns	671	70 (10.4)	173 (25.8)
MOBA3	newborns	252	28 (11.1)	73 (29.0)
NEST	newborns	413	69 (16.7)	136 (32.9)
NFCS	newborns	889	245 (27.6)	325 (36.6)
VIVA	newborns	485	14 (2.9)[b]	26 (5.4)
ALSPAC	older children	840	89 (10.6)	115 (13.7)
BAMSE	older children	347	26 (7.5)	43 (12.4)
GALA II	older children	569	40 (7.0)	76 (13.4)
MeDALL	older children	851	86 (10.2)	121 (14.3)
SEED	older children	584	25 (4.3)	49 (8.4)

[a] N=number of participants with smoking data, 450K methylation, and covariates. Participants who quit smoking during pregnancy were not included in the sustained smoking models.

[b] Cohorts where the sustained smoking category had N<15 or insufficient information to create the requested category, resulting in exclusion from the sustained smoking analysis models. All cohorts were included in the models evaluating the exposure any smoking during pregnancy.

Table 2. Meta-analysis results from newborns for selected loci not previously reported with genome-wide statistically significant differential methylation in newborn DNA in relation to sustained maternal smoking in pregnancy: Selection limited to genes prioritized for discussion.

CHR	Position	CpG	Mapped Gene	Nearest Gene (10 Mb)	Gene Group	Coef	SE	P	Direction of effect across cohorts	Mean Beta
1	24648203	cg06376426	*GRHL3*	*GRHL3*	TSS1500;Body	-0.004	0.001	1.84E-04	-+-------	0.262
2	43685377	cg20629315	*THADA*	*THADA*	Body	0.003	0.001	4.62E-04	--+++++++	0.896
2	206628553	cg22308949	*NRP2*	*NRP2*	Body	-0.016	0.002	7.83E-12	---------	0.413
2	206628625	cg05348875	*NRP2*	*NRP2*	Body	-0.026	0.004	1.13E-10	---------	0.613
2	206628692	cg14157435	*NRP2*	*NRP2*	Body	-0.028	0.004	1.61E-10	---------	0.413
2	206692685	cg14400541	-	*NRP2*	-	-0.008	0.002	5.19E-05	+--------	0.501
3	189348936	cg05129081	*TP63*	*TP63*	TSS1500	0.012	0.002	1.21E-07	++++++-++	0.539
3	189349021	cg06720722	*TP63*	*TP63*	TSS200	0.009	0.002	8.49E-06	++++++-++	0.798
4	10117479	cg22821355	*WDR1*	*WDR1*	Body	-0.007	0.002	1.50E-04	---------	0.382
4	81109888	cg01789499	*PRDM8*	*PRDM8*	5'UTR	-0.009	0.002	1.86E-04	------+--	0.852
4	81110205	cg09595050	*PRDM8*	*PRDM8*	5'UTR	-0.018	0.004	1.71E-06	---------	0.739
4	81110459	cg14197071	*PRDM8*	*PRDM8*	5'UTR	-0.021	0.005	5.33E-06	------+--	0.723
4	81111177	cg27111250	*PRDM8*	*PRDM8*	5'UTR	-0.020	0.005	1.61E-05	------+--	0.708
4	81111393	cg27639662	*PRDM8*	*PRDM8*	5'UTR	-0.016	0.004	3.33E-05	------+--	0.702
4	81117647	cg05452645	*PRDM8*	*PRDM8*	TSS1500;5'UTR	-0.022	0.004	8.95E-09	------+--	0.520
4	81117665	cg00138041	*PRDM8*	*PRDM8*	TSS1500;5'UTR	-0.021	0.004	1.39E-06	------+--	0.556
4	81117853	cg06373870	*PRDM8*	*PRDM8*	TSS1500;5'UTR	-0.017	0.003	1.24E-08	------+--	0.422
4	81118188	cg03463411	*PRDM8*	*PRDM8*	TSS1500;5'UTR	-0.014	0.003	1.09E-06	------+--	0.372
4	81118343	cg04235768	*PRDM8*	*PRDM8*	TSS1500;5'UTR	-0.014	0.002	1.35E-09	------+--	0.159
4	81118588	cg26299084	*PRDM8*	*PRDM8*	5'UTR;TSS200	-0.012	0.003	1.98E-06	------+--	0.247
4	81118794	cg06307913	*PRDM8*	*PRDM8*	5'UTR;1stExon	-0.020	0.003	3.72E-09	------+--	0.424
4	81119178	cg27242132	*PRDM8*	*PRDM8*	5'UTR	-0.022	0.004	2.98E-09	------+--	0.240
4	81119198	cg18073471	*PRDM8*	*PRDM8*	5'UTR	-0.018	0.003	1.21E-08	------+--	0.178
4	81119249	cg02458885	*PRDM8*	*PRDM8*	5'UTR	-0.010	0.002	5.94E-06	------+--	0.189
4	81119299	cg11388320	*PRDM8*	*PRDM8*	5'UTR	-0.023	0.004	1.12E-08	---------	0.324
4	81119473	cg22902505	*PRDM8*	*PRDM8*	5'UTR	-0.027	0.004	1.21E-10	------+--	0.433
4	81122726	cg05522011	*PRDM8*	*PRDM8*	Body	-0.015	0.004	2.00E-04	------+--	0.799
5	78365647	cg01856645	*DMGDH;BHMT2*	*BHMT2*	TSS200;Body	0.008	0.002	3.35E-06	++++++-++	0.177
5	78365687	cg06501366	*BHMT2;DMGDH*	*BHMT2*	Body;TSS1500	0.018	0.003	1.11E-10	+++++++++	0.408
5	78365691	cg08328513	*BHMT2;DMGDH*	*BHMT2*	Body;TSS1500	0.017	0.003	3.94E-09	++++++-++	0.265
5	78365710	cg23911707	*BHMT2;DMGDH*	*BHMT2*	Body;TSS1500	0.006	0.001	5.69E-06	+++++++++	0.260
5	78365801	cg03400060	*BHMT2;DMGDH*	*BHMT2*	Body;TSS1500	0.012	0.002	2.96E-10	+++++++++	0.392
5	78366076	cg01902605	*BHMT2;DMGDH*	*BHMT2*	Body;TSS1500	0.013	0.002	1.50E-09	+++++++++	0.707
6	7673306	cg25370658	-	*BMP6*	-	0.004	0.001	2.63E-04	+++-+--++	0.806
6	7698374	cg17951878	-	*BMP6*	-	0.013	0.003	1.15E-06	+++++++++	0.286
6	7731280	cg23623251	*BMP6*	*BMP6*	Body	0.006	0.002	3.25E-04	+++++++++	0.783
6	10405499	cg16199280	*TFAP2A*	*TFAP2A*	Body	0.006	0.002	3.26E-04	+-+++++++	0.342
6	55767865	cg16728651	-	*BMP5*	-	-0.010	0.002	1.05E-06	---------	0.729
6	152011415	cg08161546	*ESR1*	*ESR1*	TSS1500	0.008	0.002	3.50E-04	+-+++++++	0.709
6	152124815	cg08415493	*ESR1*	*ESR1*	5'UTR	-0.003	0.001	1.74E-04	-+-------	0.706
6	152126736	cg20893956	*ESR1*	*ESR1*	5'UTR;TSS200	-0.009	0.002	4.13E-05	--+---+--	0.620
6	152126785	cg07746998	*ESR1*	*ESR1*	5'UTR;TSS200	-0.006	0.002	1.18E-04	-+-------	0.594
6	152126895	cg21157690	*ESR1*	*ESR1*	5'UTR;1stExon	-0.008	0.002	5.70E-05	-++------	0.747
6	152126938	cg17264271	*ESR1*	*ESR1*	5'UTR;1stExon	-0.009	0.002	1.26E-06	-++------	0.627
6	152130058	cg04063345	*ESR1*	*ESR1*	Body	-0.013	0.004	1.22E-04	-++---+--	0.507
6	152130207	cg15626350	*ESR1*	*ESR1*	Body	-0.018	0.004	1.42E-06	-+----+--	0.444

CHR	Position	CpG	Mapped Gene	Nearest Gene (10 Mb)	Gene Group	Coef	SE	P	Direction of effect across cohorts	Mean Beta
6	152130332	cg00601836	*ESR1*	*ESR1*	Body	-0.014	0.003	1.19E-06	-++------	0.676
8	1403050	cg16442298	-	*DLGAP2*	-	-0.004	0.001	3.28E-04	---------	0.716
8	1404023	cg03551508	-	*DLGAP2*	-	-0.007	0.002	2.74E-05	---------	0.746
8	1427491	cg00827210	-	*DLGAP2*	-	-0.007	0.001	4.83E-07	------+--	0.882
8	1442292	cg13063207	-	*DLGAP2*	-	-0.006	0.001	1.85E-05	---------	0.847
8	1458508	cg24526596	*DLGAP2*	*DLGAP2*	5'UTR	-0.005	0.001	2.52E-04	------+--	0.590
8	1462903	cg25955692	*DLGAP2*	*DLGAP2*	5'UTR	-0.005	0.001	1.67E-05	++-------	0.874
8	1468625	cg00598912	*DLGAP2*	*DLGAP2*	5'UTR	-0.003	0.001	1.39E-04	---------	0.831
8	1494546	cg23424125	*DLGAP2*	*DLGAP2*	5'UTR	-0.010	0.003	3.23E-04	------+--	0.850
8	1501226	cg03185622	*DLGAP2*	*DLGAP2*	Body	-0.005	0.001	5.33E-07	+----+---	0.825
8	1526540	cg15833940	*DLGAP2*	*DLGAP2*	Body	-0.013	0.003	6.70E-06	---------	0.659
8	1534376	cg02840179	*DLGAP2*	*DLGAP2*	Body	-0.004	0.001	1.05E-04	---------	0.816
8	1615080	cg02709139	*DLGAP2*	*DLGAP2*	Body	-0.007	0.002	1.32E-05	------+--	0.870
8	1616381	cg04687241	*DLGAP2*	*DLGAP2*	Body	-0.008	0.002	6.92E-06	-+-------	0.666
8	1618448	cg06040034	*DLGAP2*	*DLGAP2*	Body	-0.013	0.003	2.42E-06	---------	0.619
8	1649758	cg02083412	*DLGAP2*	*DLGAP2*	3'UTR	-0.004	0.001	3.15E-05	--------+	0.129
8	1649868	cg22763586	*DLGAP2*	*DLGAP2*	3'UTR	-0.013	0.003	5.83E-07	------+--	0.450
8	1650172	cg27351978	*DLGAP2*	*DLGAP2*	3'UTR	-0.015	0.004	8.69E-05	------+--	0.566
8	1650309	cg02690013	*DLGAP2*	*DLGAP2*	3'UTR	-0.013	0.003	9.92E-06	------+--	0.599
14	54412780	cg23104439	-	*BMP4*	-	0.005	0.001	2.08E-04	++++++++-	0.741
14	54418728	cg05928290	*BMP4*	*BMP4*	Body	0.024	0.003	1.48E-19	+++++++++	0.759
14	54418804	cg05923197	*BMP4*	*BMP4*	Body	0.029	0.003	1.08E-18	+++++++++	0.699
14	54418851	cg09367901	*BMP4*	*BMP4*	Body	0.019	0.002	3.98E-17	+++++++++	0.827
14	54419614	cg08046044	*BMP4*	*BMP4*	5'UTR	0.005	0.001	2.70E-09	+++-+++++	0.077
14	54424149	cg24526899	*BMP4*	*BMP4*	TSS1500	0.007	0.002	5.14E-04	++++++++-	0.441
17	76930245	cg04999637	-	*TIMP2*	-	0.005	0.001	6.91E-05	+++++++++	0.583

[a] Meta-analysis results of the association between sustained maternal smoking during pregnancy and DNA methylation in newborns, adjusted for covariates, using normalized methylation betas as the outcome. Selected not previously reported loci genome wide significant after FDR correction. Results sorted by the chromosome and position of the CpG sites listed. Column headers: CHR: chromosome; Mapped Gene: UCSC annotated gene; Nearest Gene: Nearest gene (within 10 Mb) symbol using Snipper software; Gene Group: UCSC gene region feature category; Coef: regression coefficient; SE: standard error; P: p value; Direction: Direction of effect across cohorts included in the statistical model: maternal smoking during pregnancy associated with increased (+) or decreased (-) methylation, or missing result (?), in alphabetical order of cohorts; Mean beta: Average of the mean methylation beta values across the newborn cohorts. For complete listing of CpGs differentially methylated in relation to sustained maternal smoking during pregnancy and for results from meta-analysis models unadjusted for covariates and adjusted for covariates and cell type see Supplementary Table 3.