Rotroff et al. BMC Genomics (2016) 17:976
DOI 10.1186/s12864-016-3310-1

BMC Genomics

RESEARCH ARTICLE Open Access

Maternal smoking impacts key biological pathways in newborns through epigenetic modification *in Utero*

Daniel M. Rotroff[1,2], Bonnie R. Joubert[3], Skylar W. Marvel[1], Siri E. Håberg[4], Michael C. Wu[5], Roy M. Nilsen[6], Per M. Ueland[7,8], Wenche Nystad[4], Stephanie J. London[3*] and Alison Motsinger-Reif[1,2,9]

Abstract

Background: Children exposed to maternal smoking during pregnancy exhibit increased risk for many adverse health effects. Maternal smoking influences methylation in newborns at specific CpG sites (CpGs). Here, we extend evaluation of individual CpGs to gene-level and pathway-level analyses among 1062 participants in the Norwegian Mother and Child Cohort Study (MoBa) using the Illumina 450 K platform to measure methylation in newborn DNA and maternal smoking in pregnancy, assessed using the biomarker, plasma cotinine. We used novel implementations of bioinformatics tools to collapse epigenome-wide methylation data into gene- and pathway-level effects to test whether exposure to maternal smoking *in utero* differentially methylated CpGs in genes enriched in biologic pathways. Unlike most pathway analysis applications, our approach allows replication in an independent cohort.

Results: Data on 485,577 CpGs, mapping to a total of 20,199 genes, were used to create gene scores that were tested for association with maternal plasma cotinine levels using Sequence Kernel Association Test (SKAT), and 15 genes were found to be associated ($q < 0.25$). Six of these 15 genes (*GFI1*, *MYO1G*, *CYP1A1*, *RUNX1*, *LCTL*, and *AHRR)* contained individual CpGs that were differentially methylated with regards to cotinine levels ($p < 1.06 \times 10^{-7}$). Nine of the 15 genes (*FCRLA*, *MIR641*, *SLC25A24, TRAK1*, *C1orf180*, *ITLN2*, *GLIS1*, *LRFN1*, and *MIR451)* were associated with cotinine at the gene-level ($q < 0.25$) but had no genome-wide significant individual CpGs ($p > 1.06 \times 10^{-7}$). Pathway analyses using gene scores resulted in 51 significantly associated pathways, which we tested for replication in an independent cohort ($q < 0.05$). Of those 32 replicated in an independent cohort, which clustered into six groups. The largest cluster consisted of pathways related to cancer, cell cycle, ERα receptor signaling, and angiogenesis. The second cluster, organized into five smaller pathway groups, related to immune system function, such as T-cell regulation and other white blood cell related pathways.

Conclusions: Here we use novel implementations of bioinformatics tools to determine biological pathways impacted through epigenetic changes *in utero* by maternal smoking in 1062 participants in the MoBa, and successfully replicate these findings in an independent cohort. The results provide new insight into biological mechanisms that may contribute to adverse health effects from exposure to tobacco smoke *in utero*.

Keywords: Smoking, Epigenetics, Pathway analysis, Cancer, *In utero*

* Correspondence: London2@niehs.nih.gov
[3]Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, PO Box 12233, MD A3-05, Research Triangle Park, NC 27709, USA
Full list of author information is available at the end of the article

Background

Although many adverse effects of maternal smoking on offspring have been well identified, little is known about the underlying biological mechanisms. [1, 2] One proposed mechanism for how *in utero* exposure to tobacco smoke may impact health is through epigenetic effects including DNA methylation. Previously, Joubert et al. collected genome-wide methylation data from 1062 MoBa mother-offspring pairs and demonstrated that maternal smoking, assessed objectively by cotinine levels, is significantly associated with 1) differential DNA methylation in genes involved in metabolism of tobacco smoke compounds, and 2) novel genes involved in diverse developmental processes not previously linked to tobacco response [3]. These findings have since been widely replicated [3–6].

It has been recognized that genome wide association studies, using single nucleotide polymorphisms, that rely on single locus variation explain little of the overall heritability of complex traits [7, 8]. While there are many potential sources of this "missing heritability", single locus analysis typically ignores a large number of loci with moderate effects, due to stringent significance thresholds. Gene-based association analysis takes a gene as basic unit for association analysis. As this method can combine genetic information given by all the markers in a gene, it can obtain more informative results and increase the capability of finding novel genes and gene sets. This method has been used as a novel complement method for SNP-based GWAS in identifying disease susceptibility genes [9, 10], and we extend such an approach to methylation data here.

Additionally, To investigate the biological processes (i.e. pathways) impacted by maternal smoking during pregnancy and associated altered fetal methylation, we performed gene set/pathway analysis to further dissect the biological impact of maternal smoking. We applied a novel approach that combines analysis tools for collapsing epigenome-wide methylation data into gene- and pathway-based effects (Fig. 1). Pathway analysis combines significant genes into sets of genes, or pathways, that are thought to have coordinated effects on a biological endpoint.

A number of pathway analysis methods have been developed, and have been widely applied in human genetics and genomics. The majority of pathway analysis methods were originally developed for microarray, gene expression data, and the most popular methods perform enrichment analysis for gene sets defined by external knowledge bases [11]. In the current study, we modified the bioinformatics approaches that have been developed in other contexts to be valid for epigenome-wide data analysis.

Importantly, we performed a two stage study, performing both discovery and replication of the gene-based and pathway-based associations. While replication is standard in genetic association studies for individual variants it is rarely performed for pathway analyses. Whether due to the limited availability of proper validation cohorts in many studies, or challenges in adapting pathway approaches to allow for a discovery and replication approach, this lack of replication is an important limitation of many pathway analysis studies. The previously described MoBa cohort, referred to as MoBa1 was used as

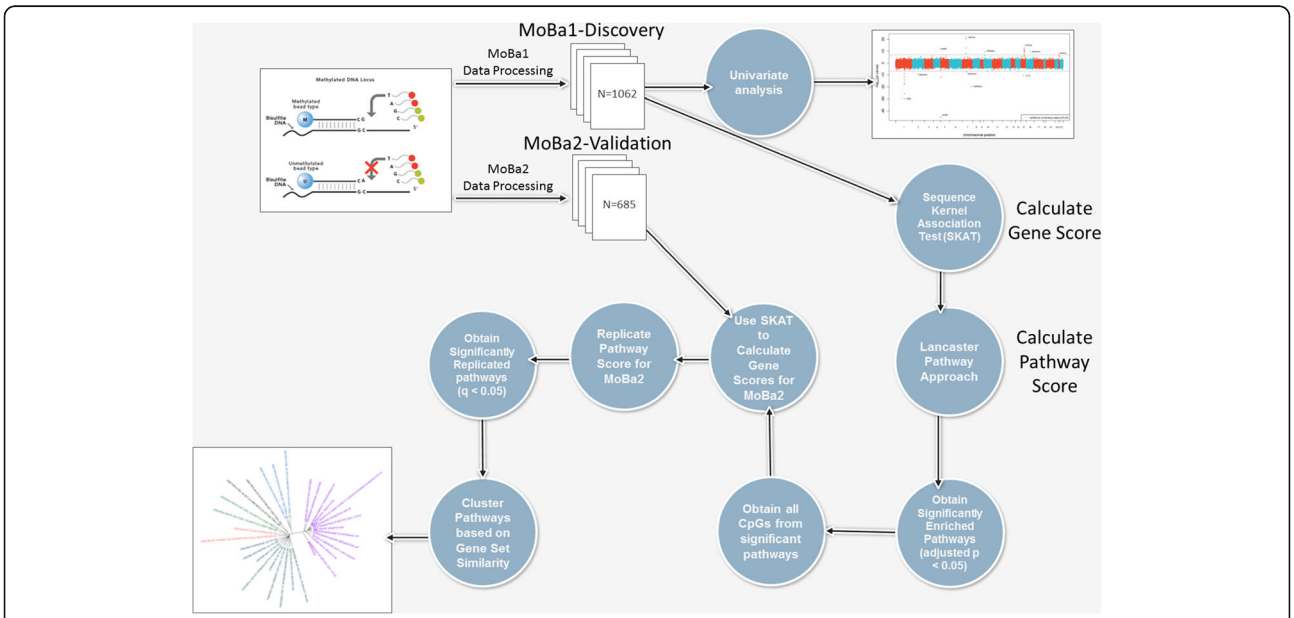

Fig. 1 Analysis workflow collapsing individual CpG data into gene- and pathway-level scores, and replication of findings

the discovery cohort. We subsequently measured DNA methylation in an additional 685 MoBa newborns; this dataset is referred to as MoBa2 and is used as the replication cohort.

Results

In univariate analysis of individual CpGs in the discovery cohort MoBa1, we found methylation at 27 CpGs in newborns to be significantly associated with maternal plasma cotinine levels analyzed as a continuous variable (Bonferroni correction for 473,864 tests, $p < 1.06 \times 10^{-7}$). The majority of those markers are annotated within genes. Twenty four markers are annotated within the *GFI1*, *AHRR*, *MYO1G*, *CNTNAP2*, *FRMD4A*, *LCTL*, *CYP1A1*, and *RUNX1* genes (Fig. 2). The three significant markers (cg00253658, cg18703066, cg04598670) that did not map to known genes are located on chr16 at 54210496, chr2 at 105363536, and chr7 at 68697651.

We then grouped individual CpGs by gene to form a gene-level *p* value, or gene score, using the Sequence Kernel Association Test (SKAT) software implemented in R [12, 13]. A total of 20,199 genes were tested and 15 were associated with maternal plasma cotinine levels with an FDR-adjusted $q < 0.25$ (Table 1). Six of these 15 genes (*GFI1*, *MYO1G*, *CYP1A1*, *RUNX1*, *LCTL*, and *AHRR*) contained genome-wide significant individual CpGs ($p < 1.06 \times 10^{-7}$). Nine of the 15 genes (*FCRLA*, *MIR641*, *SLC25A24*, *TRAK1*, *C1orf180*, *ITLN2*, *GLIS1*, *LRFN1*, and *MIR451*) were associated with cotinine ($q < 0.25$) but did not have any genome-wide significant individual CpGs (Table 1). This demonstrates the utility of this method to detect important effects at a gene-level that would have otherwise gone undetected by interrogating only individual CpGs.

Only two genes, *CNTNAP2* and *FMRD4A*, had genome-wide significant individual CpGs ($p < 1.06 \times 10^{-7}$), but did not result in gene scores with $q < 0.25$. Eighty CpGs mapped to *CNTNAP2* but only one (cg25949550), located in the gene body, was statistically significant ($q = 1.07 \times 10^{-13}$) resulting in a gene score ($q = 0.32$) that did not reach our threshold for association (Additional file 1). There were 127 CpGs mapped to *FRMD4A* on this platform and only two CpGs (cg11813497, cg15507334), located within 200 bp of the transcriptional start site, were at or near genome-wide significance, for an overall gene score with a $q = 0.28$ (Additional file 1).

We then collapsed the gene-level results into pathway level statistics using *a priori* pathway gene sets from the MSigDB database. MSigDB provides annoted collections of gene sets curated from multiple biological knowledgebases. We selected relevant gene sets as described below to collapse individual gene association scores into pathway analysis results. A total of 5836 pathway gene sets were tested for association using a the correlated Lancaster *p*-value approach. After a Bonferroni correction ($p < 0.05$) for the number of pathways tested, a total of 51 pathways were statistically significant in the (Fig. 1 and Table 2). Pathways spanned a range of physiological

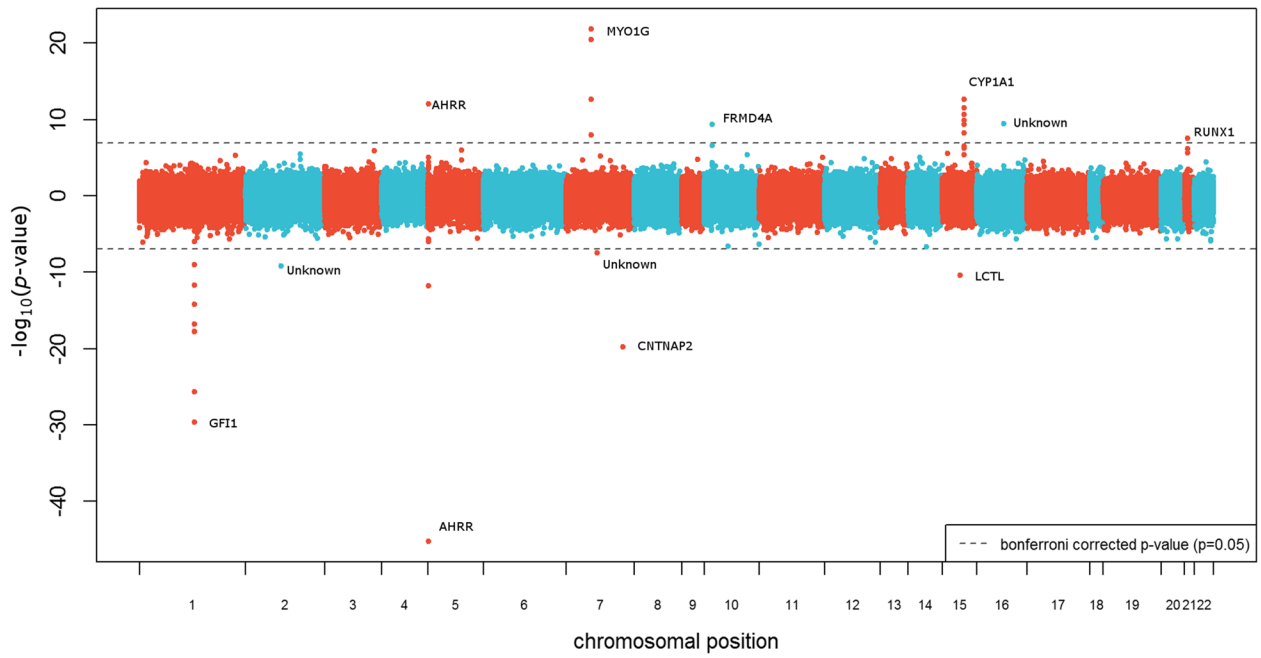

Fig. 2 Manhattan plot of univariate CpG results. The y-axis represents the –log10 of the CpG *p*-values. CpGs with negative *p*-values corresponded to decreased methylation, whereas positive *p*-values corresponded to increased methylation. CpGs that reached genome-wide significance, with a bonferonni corrected $p < 0.05$ are annotated with their corresponding genes

Table 1 Genes differentially methylated in newborns in relation to maternal smoking during pregnancy using the Sequence Kernel Association Test (SKAT) in the MoBa1 discovery cohort ($n = 1062$ subjects)

Gene[a]	Markers/Gene	SKAT *p*-value	SKAT *q*-value
GFI1	71	1.05E-17	2.13E-13
MYO1G	12	4.33E-17	4.37E-13
CYP1A1	35	1.21E-09	8.15E-06
RUNX1	53	3.46E-07	0.001749
LCTL	8	1.61E-05	0.065098
AHRR	149	6.29E-05	0.184672
FCRLA	9	8.14E-05	0.184672
MIR641	4	8.23E-05	0.184672
TRAK1	35	7.78E-05	0.184672
C1ORF180	4	0.000104	0.209611
ITLN2	5	0.000116	0.212334
GLIS1	51	0.000156	0.223673
LRFN1	21	0.00016	0.223673
MIR451	8	0.000166	0.223673
SLC25A24	23	0.000144	0.223673

[a] Covariates included: maternal education, CD8T, CD4T, natural killer cell fraction, B cell fraction, monocyte fraction, granulocyte fraction

and pathophysiological functions including cell cycle, cancer, white blood cell differentiation, genotoxicity, and others (Additional file 2).

Subsequently, we attempted to replicate the pathway analysis by calculating gene scores in the MoBa2 replication cohort data for all genes in the 51 statistically significant pathways from the MoBa1 discovery cohort. Gene and pathway level association scores were calculated identically to the procedure described for the discovery cohort (Fig. 1), and a FDR correction was used to correct for multiple testing. Of the 51 pathways identified in the MoBa1 cohort ($p < 8.6 \times 10^{-6}$), 32 replicated ($q < 0.05$) (Table 2).

Because of the relatively large number of pathways that replicated across both cohorts, we performed clustering analysis to aid in interpretability. We clustered replicated pathways according to gene set similarity (Fig. 3). We identified six clusters, or groups, of pathways that contained similar gene sets and were reflective of their biological function. The largest cluster consisted of pathways related to cancer (FALVELLA SMOKERS WITH LUNG CANCER, HEDENFALK BREAST CANCER BRACX UP), cell cycle (INTERPHASE OF MITOTIC CELL CYCLE, INTERPHASE, G1 S TRANSITION OF MITOTIC CELL CYCLE), ERα receptor signaling (WILLIAMS ESR1 TARGETS DN, FRASOR RESPONSE TO ESTRADIOL UP), and angiogenesis (ABE VEGFA TARGETS 2HR, ELVIDGE HIF1A TARGETS DN). A second cluster was organized into five smaller pathway groups related to immune system function, such as T-cell regulation (e.g. GSE1460 DP THYMOCYTE VS NAIVE CD4 TCELL ADULT BLOOD UP, GSE3982 DC VS TH1 DN, GSE3982 CENT MEMORY CD4 TCELL VS TH1 DN) and other white blood cell related pathways (e.g. GSE1460 DP VS CD4 THYMOCYTE UP, CASORELLI ACUTE PROMYELOCYTIC LEUKEMIA UP).

Discussion

There is an overwhelming body of epidemiological evidence linking smoking during pregnancy to various health outcomes in the offspring including low birth weight, reduced lung function, and increased respiratory infections [1]. Additional associations have also been reported between maternal smoking during pregnancy and 1) rheumatoid arthritis and other inflammatory polyarthropathies [14–17], 2) child behavior and cognitive functioning, and 3) mixed results of associations with childhood cancers. While these associations are consistent, the underlying mechanisms leading to these outcomes have remained elusive. The analyses presented here support the possibility that epigenetic mechanisms may play a role, and point towards a number of pathways that may be involved.

Multiple pathways related to T-cell function were altered by maternal smoking. GFI1, previously reported by Joubert et al. [3], was a main driver for many of the T-cell, eosinophil, and neutrophil related pathway scores (e.g. GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_UP, GSE3982_CENT_MEMORY_CD4_TCELL_VS_TH1_DN, GSE3982_NEUTROPHIL_VS_TH1_DN, GSE3982_EOSINOPHIL_VS_TH1_DN). Additional genes that contributed to the impact on immune response pathways include *IL22* ($p = 0.039$, $q = 0.28$) and *IL2RA* ($p = 0.002$, $q = 0.28$) which were not detected in the analysis of Joubert et al. [3] based on single CpGs.

IL22 is a cytokine involved in the initiation of innate immune response against pathogens, and is especially active in epithelial cells of the gut and lung [18]. Reduced expression of *IL2RA* on the surface of immune cells has been known to cause chronic immune suppression and may be linked to type 1 diabetes mellitus [19, 20]. Collectively, these pathways are relevant to various health effects in newborns that have been associated with exposure to maternal smoking during pregnancy [14, 17, 21].

Mixed results have been found regarding in utero tobacco exposure and increased incidence of childhood cancers. Some studies have found increased risk of childhood cancers with maternal smoking during pregnancy [16, 22], whereas, others have found null results [15, 23]. However, here we present evidence that alterations in methylation may affect key pathways related to cancer.

Table 2 Significantly enriched pathways based on differential methylation in newborns exposed to maternal smoking during pregnancy

Pathway Name	MSigDB Contributor[a]	MSigDB Category Code	# Genes Pathway	# Genes Overlap	Discovery *p* value	Bonferroni Adjusted Discovery *p* value	Replication *p* value	Replication *q* value	Bonferroni Adjusted Replication *p* value
GSE17974_0H_VS_12H_IN_VITRO_ACT_CD4_TCELL_UP	Nick Haining lab (DFCI)	C7	200	172	1.56E-14	9.12E-11	9.31E-07	2.60E-05	4.28E-05
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_2H_CD4_TCELL_UP	Nick Haining lab (DFCI)	C7	200	171	3.01E-14	1.76E-10	1.13E-06	2.60E-05	5.19E-05
GSE17974_0H_VS_6H_IN_VITRO_ACT_CD4_TCELL_UP	Nick Haining lab (DFCI)	C7	200	169	1.55E-17	9.07E-14	5.33E-06	8.17E-05	0.0002
G1_S_TRANSITION_OF_MITOTIC_CELL_CYCLE	GO	C5	27	27	1.16E-18	6.74E-15	0.0012	0.0135	0.0548
WILLIAMS_ESR1_TARGETS_DN	Broad Institute	C2	6	6	1.83E-08	0.000107	0.0015	0.0135	0.0675
TIEN_INTESTINE_PROBIOTICS_2HR_UP	Broad Institute	C2	27	26	2.54E-08	0.000148	0.0019	0.0144	0.0862
TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_SUSTAINED_IN_MONOCYTE_UP	Broad Institute	C2	21	21	6.48E-09	3.78E-05	0.0027	0.0158	0.1226
INTERPHASE_OF_MITOTIC_CELL_CYCLE	GO	C5	62	59	1.17E-16	6.85E-13	0.0027	0.0158	0.1264
HEDENFALK_BREAST_CANCER_BRACX_UP	University of Washington	C2	20	14	4.81E-10	2.80E-06	0.0032	0.0159	0.1464
ABE_VEGFA_TARGETS_2HR	University of Washington	C2	34	30	2.16E-09	1.26E-05	0.0036	0.0159	0.1675
INTERPHASE	GO	C5	68	65	7.20E-17	4.20E-13	0.0038	0.0159	0.1753
FRASOR_RESPONSE_TO_ESTRADIOL_UP	Broad Institute	C2	37	37	3.72E-08	0.000217	0.0046	0.0175	0.2099
ENGELMANN_CANCER_PROGENITORS_UP	Broad Institute	C2	48	47	4.18E-07	0.002441	0.0059	0.0209	0.2713
MIKKELSEN_IPS_WITH_HCP_H3K27ME3	Broad Institute	C2	102	97	4.80E-13	2.80E-09	0.0073	0.0239	0.3346
FALVELLA_SMOKERS_WITH_LUNG_CANCER	Broad Institute	C2	80	71	5.90E-08	0.000344	0.0091	0.0278	0.4165
AMUNDSON_GENOTOXIC_SIGNATURE	Broad Institute	C2	105	94	1.22E-15	7.15E-12	0.0110	0.0305	0.5045
GSE1460_DP_VS_CD4_THYMOCYTE_UP	Nick Haining lab (DFCI)	C7	200	174	1.29E-16	7.53E-13	0.0113	0.0305	0.5186
GSE3982_DC_VS_TH1_DN	Nick Haining lab (DFCI)	C7	200	174	4.75E-18	2.77E-14	0.0124	0.0307	0.5702
ELVIDGE_HIF1A_TARGETS_DN	Broad Institute	C2	91	85	2.82E-08	0.000165	0.0127	0.0307	0.5824
CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_UP	Broad Institute	C2	177	150	9.14E-14	5.34E-10	0.0140	0.0322	0.6444

Table 2 Significantly enriched pathways based on differential methylation in newborns exposed to maternal smoking during pregnancy *(Continued)*

GSE1460_DP_THYMOCYTE_VS_NAIVE_CD4_TCELL_ADULT_BLOOD_UP	Nick Haining lab (DFCI)	C7	200	170	3.93E-19	2.29E-15	0.0164	0.0353	0.7547
GSE17974_CTRL_VS_ACT_IL4_AND_ANTI_IL12_2H_CD4_TCELL_DN	Nick Haining lab (DFCI)	C7	200	181	1.27E-16	7.42E-13	0.0172	0.0353	0.7901
GSE22886_NAIVE_CD4_TCELL_VS_48H_ACT_TH2_DN	Nick Haining lab (DFCI)	C7	200	183	1.11E-15	6.48E-12	0.0194	0.0353	0.8933
GSE24634_NAIVE_CD4_TCELL_VS_DAY10_IL4_CONV_TREG_DN	Nick Haining lab (DFCI)	C7	200	185	5.91E-15	3.45E-11	0.0195	0.0353	0.8948
GSE3982_CENT_MEMORY_CD4_TCELL_VS_TH1_DN	Nick Haining lab (DFCI)	C7	200	185	1.44E-14	8.38E-11	0.0195	0.0353	0.8973
GSE17974_0H_VS_4H_IN_VITRO_ACT_CD4_TCELL_DN	Nick Haining lab (DFCI)	C7	200	182	1.43E-14	8.32E-11	0.0205	0.0353	0.9451
GSE3982_EOSINOPHIL_VS_TH1_DN	Nick Haining lab (DFCI)	C7	200	189	7.96E-15	4.65E-11	0.0209	0.0353	0.9599
GSE3982_NEUTROPHIL_VS_TH1_DN	Nick Haining lab (DFCI)	C7	200	182	2.96E-16	1.73E-12	0.0215	0.0353	0.9897
GSE24634_NAIVE_CD4_TCELL_VS_DAY7_IL4_CONV_TREG_DN	Nick Haining lab (DFCI)	C7	200	189	1.20E-15	7.00E-12	0.0250	0.0388	1
GSE15215_CD2_POS_VS_NEG_PDC_DN	Nick Haining lab (DFCI)	C7	200	180	3.26E-17	1.90E-13	0.0253	0.0388	1
GSE10856_CTRL_VS_TNFRSF6B_IN_MACROPHAGE_DN	Nick Haining lab (DFCI)	C7	200	170	1.47E-06	0.008573	0.0274	0.0407	1
GSE3982_MAC_VS_TH2_DN	Nick Haining lab (DFCI)	C7	200	182	5.98E-08	0.000349	0.0296	0.0426	1

[a] Contributor to the corresponding pathway in MSigDB. Additional information about these contributors can be found at: http://www.broadinstitute.org/gsea/msigdb/collection_details.jsp

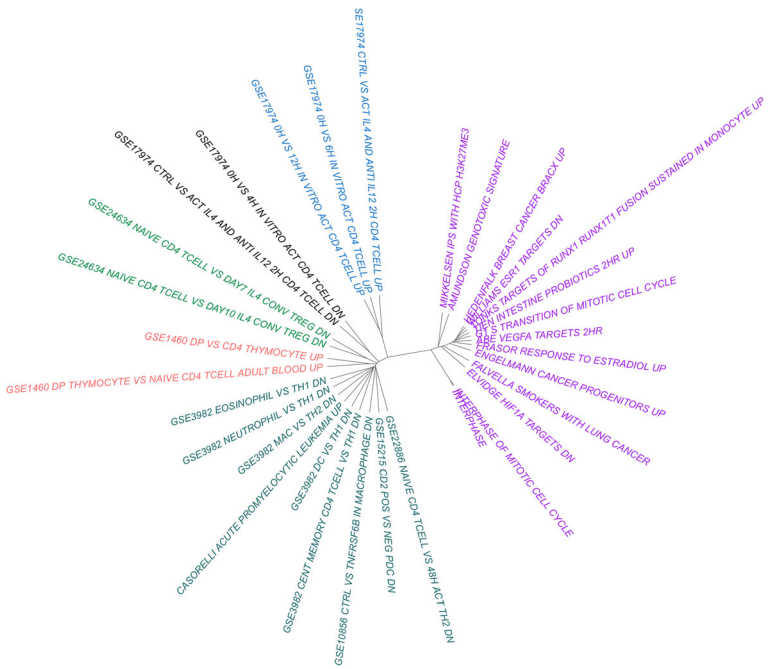

Fig. 3 Hierarchical clustering of replicated pathways. Replicated pathways were measured for similarity and clustered based on overlapping genes between each pathway. The dendrogram was cut to show six distinct clusters; pathways within the same cluster are annotated with matching colors

Joubert et al. [24] demonstrated that maternal smoking affects newborn methylation if the mother smokes through gestational week 18, whereas significant effects on methylation were not observed for mothers that quit before 18 gestational weeks. Some studies assessed smoking during pregnancy as any smoking versus no smoking. Thus if sustained smoking during pregnancy is required, as suggested by the methylation analyses, associations with cancer might be attenuated or missed entirely.

In addition to cancer-specific pathways (i.e. HEDENFALK_BREAST_CANCER_BRACX_UP, ENGELMANN_CANCER_PROGENITORS_UP, FALVELLA_SMOKERS_WITH_LUNG_CANCER, CASORELLI_ACUTE_PROMYELOCYTIC_LEUKEMIA_UP), changes in pathways related to cell cycle were detected, which are also relevant to cancer (i.e. G1_S_TRANSITION_OF_MITOTIC_CELL_CYCLE, INTERPHASE_OF_MITOTIC_CELL_CYCLE). These pathway level effects were also mainly driven by *GFI1*.

However, decreased methylation of the gene Speedy (*SPDYA*) ($p = 0.024$, $q = 0.28$) also contributed to the impact on INTERPHASE_OF_MITOTIC_CELL_CYCLE. *SPDYA* was not identified in the analysis of individual CpGs by Joubert et al. [3]. It is a cell cycle regulator that has been shown to increase cell proliferation through activation of cyclin dependent kinase-2 (cdk2) during the G1/S phase of cellular replication [25]. The ABE_VEGFA_TARGETS_2HR pathway, related to vascular endothelial growth factor-A gene (*VEGFA*), was significantly altered (replication $q = 0.03$). *VEGFA* mediates angiogenesis, suppresses apoptosis, and is the pharmacological target for Bevacizumab, a monoclonal antibody chemotherapeutic drug [26–28]. *VEGFA* is increased during oxidative stress and results in a compensatory increase in angiogenesis, a hallmark of cancer [28–30].

Furthermore, impacts on pathways WILLIAMS_ESR1_TARGETS_DN and FRASOR_RESPONSE_TO_ESTRADIOL_UP point towards effects related to estrogen receptor-alpha (ERα) signaling which is important in several cancers [31–33]. Effects on these pathways were largely mediated through CYP1A1 ($p = 1.21 \times 10^{-9}$), which was previously identified by Joubert et al., and PDZK1 ($p = 0.0007$) which was not.

Effects on pathways related to cell cycle and angiogenesis may also point towards mechanisms by which birth weight may be affected. Recently, a study by Miller et al. [34] demonstrated a differential effect on male birth weight from non-smoking mothers if the maternal grandmother smoked while pregnant, suggesting a potential epigenetic mechanism may be responsible. Decreased birth weight is a well-established effect of maternal smoking on offspring, although the mechanism by which this occurs has not been elucidated [35].

Through the novel implementation of methods for creating gene scores [13] and pathway scores [36], we

have identified and replicated key biological processes related to maternal smoking via its impact on newborn DNA methylation. These methods permit replication, which limits the likelihood of false-positive findings. To our knowledge, until now no studies of pathway impacts on methylation have been performed in tandem with a replication dataset. Furthermore, using gene based tests, we identified associations with genes not identified by CpG specific analyses alone – these included *FCRLA*, *MIR641*, *SLC25A24*, *TRAK1*, *C1orf180*, *ITLN2*, *GLIS1*, *LRFN1*, and *MIR451*.

The replicated pathway analysis conducted offers potential new insights into the biological impacts of maternal smoking on fetal DNA methylation. The genes and pathways detected point to effects on T-cell mediation, cell cycle, and xenobiotic metabolism. In turn, these data further support a potential epigenetic role for the adverse health effects observed in children exposed to maternal smoking during pregnancy.

Methods

Study population

Participants in this analysis include 1062 mother-offspring pairs from a substudy of the Norwegian Mother and Child Cohort Study (MoBa) [37–39]. In a previous study with this cohort, individual CpG sites in newborns were tested for differential methylation in relation to maternal smoking [3]. This dataset is referred to as MoBa1 and was used as the discovery cohort. We subsequently measured DNA methylation in an additional 685 newborns. This dataset is referred to as MoBa2 and was used as the replication cohort. The study has been approved by the Regional Committee for Ethics in Medical Research, the Norwegian Data Inspectorate and the Institutional Review Board of the National Institute of Environmental Health Sciences, USA, and written informed consent was provided by all mothers participating.

Covariates and cotinine measurements

Information on maternal age, parity, and maternal education was collected from 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 Joubert et al. [3], indicative of less than high school/secondary school, high school/secondary school completion, some college or university, and 4 years of college/university or more. Maternal smoking during pregnancy (none, stopped before 18 weeks of pregnancy, smoked past 18 weeks of pregnancy) was assessed by maternal questionnaire and verified with maternal plasma cotinine measured by liquid chromatography - tandem mass spectrometry at approximately 18 weeks gestation [40].

For MoBa1, cotinine, a quantitative biomarker of smoking, was measured in maternal plasma and was analyzed as a continuous variable. No cotinine was detected in 736 participants, and of the participants with detectable cotinine levels ($N = 326$) the mean cotinine level was 191 (SE = 11). For MoBa2, cotinine measurements were not available for most mothers. Therefore, a three-category variable based on the mother's report of smoking during pregnancy was created and supported using cotinine measurements when available ($N = 221$ MoBa2 participants had cotinine data). The three categories represented no smoking ($N = 512$), stopped during pregnancy ($N = 103$), or smoked throughout pregnancy ($N = 70$).

Methylation measurements

Details of the DNA methylation measurements and quality control for the MoBa1 participants were previously described [3] and the same reagents, platforms and protocols were used for the MoBa2 participants. All biological material was obtained from the Biobank of the MoBa study [38]. Briefly, DNA was extracted from umbilical cord whole blood samples [36]. 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 [41, 42]. The package *minfi* in R was used 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)) from the raw intensity (idat) files [43, 44].

Probe and sample-specific quality control filtering was performed separately in MoBa1 and MoBa2 datasets. 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 mismatches ($N = 13$ MoBa1, $N = 8$ MoBa2) were also removed. For each dataset, we accounted for the two different probe designs by applying the intra-array normalization strategy Beta Mixture Quantile dilation (BMIQ) [45].

The gPCA program was used to determine the presence of batch effects, using plate to represent batch and ComBat was applied for batch correction using the SVA package in R for both MoBa 1 and MoBa 2 cohorts [44, 46–48]. A total of 473,772 markers remained

after data processing, and 365,860 of these markers mapped to at least one of 21,231 genes using Illumina provided annotation based on human reference genome [NCBI build 37].

Covariate selection

All analysis was conducted in the statistical programming language, R [44]. Initially, potential clinical and demographic variables: maternal age, newborn gender, education, asthma, folate, and parity were evaluated as potential covariates prior to association analysis. Each potential covariate was tested for association with maternal cotinine using linear least squares regression, with categorical variables dummy encoded in the model(s). Two-sided *p*-values from each regression analysis were recorded, and a False Discovery Rate (FDR) correction for multiple comparisons was applied to limit false positives. Covariates with an FDR-adjusted q value < 0.1 were included in subsequent models [49]. In addition, cell type fractions (CD8T, CD4T, natural killer cell, B cell, monocyte, granulocyte) for each subject were calculated using the reference-based Houseman method in the *minfi* package in *R* [43, 44, 50], and these fractions were forced as covariates into subsequent models. The same selection criteria was used for both the discovery and replication dataset. The only resulting covariate was maternal education for MoBa1 ($q < 0.1$), and maternal age, education, folate, and parity were selected as covariates for MoBa2 ($q < 0.1$).

Univariate association analysis

Statistical tests for the association of each CpG marker and maternal plasma cotinine levels (continuous) were performed using linear least-squares regression for the MoBa1 cohort. Significant covariates and cell type fractions were included in the model to reduce confounding. All CpG *p* values, on the $-\log_{10}$ scale, were plotted according to genomic sequence in a Manhattan plot (Fig. 1).

Gene score calculation

To perform gene-level association analysis, CpG markers were collapsed by gene using the Illumina provided annotation based on human reference genome [NCBI build 37]. For each gene, the CpG data was combined into a gene-level *p* value using the Sequence Kernel Association Test (SKAT) software implemented in R [12, 13]. The SKAT null model for MoBa1 was created using significantly associated covariates: maternal education ($q < 0.1$), and cell type fractions (CD8T, CD4T, natural killer cell, B cell, monocyte, granulocyte). The same modeling strategy was implemented for the SKAT null model for MoBa2 and included significantly associated covariates and the cell type fractions. The SKAT model was then run using an unweighted, linear kernel with the 'is_check_genotype' flag set to FALSE. In order to account for the underlying correlation structure for the *p* value gene scores, the SKAT null model was created with the cotinine values and covariate values randomly shuffled, and then SKAT was run on the residuals until 1000 permuted gene scores were created. To control for multiple comparisons, we report gene scores with a FDR $q < 0.25$ as being associated with cotinine levels.

Pathway analysis

The results from the SKAT gene-level association analysis (specifically *p*-values) were used for pathway-level analysis. Genes were grouped into a priori pathways (gene sets) using the Molecular Signatures Database v4.0 (MSigDB) [51]. MSigDB contains gene sets from a collection of popular resources such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [51]. A subset of pathways was selected for analysis based on a set of four criteria: 1) the pathway must be composed of a set of genes from *Homo sapiens,* 2) the number of genes in a pathway cannot exceed 250 genes, 3) at least one gene in the pathway must be present in the list of available gene scores, and 4) pathways representing positional gene sets (C1), motif gene sets (C3), and computational derived gene sets (C4) were excluded. This resulted in a total of 5836 pathways for analysis. These pathways came from the either curated gene sets (C2), GO gene sets (C5), oncogenic signatures gene sets (C6), or the immunologic signatures gene sets (C7) collections in MSigDB [9]. Each pathway consists of a set of genes that are considered biologically relevant to a given biological function or signaling network, and individual genes are often represented in multiple pathways.

The pathway-level score was calculated from the individual gene scores that overlapped with the genes in each pathway gene set. The pathway level score is the combined *p*-value across all gene-level results from the SKAT analysis. There are a number of approaches for combining *p*-values, but most assume that the individual *p*-values are not correlated. Pathway analysis actually relies on the fact that genes scores within a pathway are correlated, so a collapsing approach that explicitly takes that into account was used. More specifically, the individual gene scores were combined into pathway-level scores using the correlated Lancaster method in Dai et al. (T_A) [36]. This resulted in a final *p*-value for each pathway from MSigDB. It is important to note that this combined *p*-value represents a self-contained pathway analysis, where the null hypothesis is that gene sets are not more strongly associated than expected by chance. Because of the large number of pathways tested, we controlled for multiple comparisons using a conservative Bonferroni correction. We chose a conservative

approach, even though the p-values from each pathway are not independent, since genes appear in multiple pathways. Pathways with a corrected $p < .05$ ($n = 5836$; $p < 8.6 \times 10^{-6}$) were considered statistically significant in the discovery cohort.

Replication

The statistically significant pathways ($p < 8.6 \times 10^{-6}$) were tested for replication using MoBa2. The CpG values were combined for genes that occurred in significant pathways in MoBa1, using SKAT as described above. Gene scores were then combined using the Lancaster approach to calculate a pathway-level score for the replication cohort. Pathways p values were adjusted using both an FDR and a more conservative Bonferroni approach and were considered to be successfully replicated with an FDR $q < 0.05$ (Table 2). Pathway analyses are commonly divided into self-contained or competitive approaches. Here we use a self-contained, global null approach to pathway analysis. An advantage of this approach is that it lends itself toward replication in smaller cohorts because only genes in significant pathways from the discovery cohort need to be tested for replication. Competitive pathway analysis methods test a different null hypothesis, and subsequently require all genes to be tested, which can make replication with smaller cohorts unfeasible.

Pathway hierarchical clustering

Hierarchical clustering was performed using R and the 'APE' package [44, 52]. All unique genes within replicated pathways ($q < .05$) were tabulated. All gene-pathway combinations were recorded as either a "1" if the pathway contained the gene or a "0" if the pathway did not contain the gene. Clustering was then performed using Euclidean distance and Ward's method. The resulting dendrogram (Fig. 3) was then cut and colored so that six groups were defined based on gene set similarity.

Conclusions

We used a novel implementation of bioinformatics tools to collapse individual CpG results to a gene score and performed pathway analysis to test for in utero epigenetic changes by maternal smoking in 1062 participants in the MoBa. By collapsing individual CpG effects to gene scores, we found significant differential methylation in 15 genes (q<0.25), nine of which were not detected by only testing individual CpGs. Furthermore, pathway analysis revealed significant associations with 51 pathways, 32 of which replicated in an independent cohort of 685 participants. Significantly associated pathways, that replicated in the independent cohort, represent diverse biological processes including cancer, cell cycle, ERα receptor signaling, angiogenesis, and immune system function. This approach may provide new insight into the biological mechanisms that may lead to adverse health effects from exposure to tobacco smoke in utero.

Additional files

Additional file 1: SKAT_GeneScor. (XLSX 1 MB)
Additional file 2: Lancaster_Pat. (XLSX 4 MB)

Abbreviations

BMIQ: Beta Mixture Quantile dilation; cdk2: cyclin dependent kinase-2; CpGs: Region where cytosine and guanine are separated by one phosphate. The cytosine at these sites can be methylated; FDR: False Discovery Rate; GO: Gene Ontology; GSEA: Gene Set Enrichment Analysis; KEGG: Kyoto Encyclopedia of Genes and Genomes; MoBa: Norwegian Mother and Child Cohort Study; MSigDB: Molecular Signatures Database v4.0; SKAT: Sequence Kernel Association Test; *SPDYA*: Speedy gene; *VEGFA*: Vascular endothelial growth factor-A gene

Acknowledgments

We are grateful to all the participating families in Norway who take part in this on-going cohort study. The authors thank Dr. Frank Day of NIEHS and Dr. Jianping Jin of Westat, Inc for expert technical assistance.

Funding

The Norwegian Mother and Child Cohort Study are supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research, NIH/NIEHS (contract no N01-ES-75558), NIH/NINDS (grant no.1 UO1 NS 047537-01 and grant no.2 UO1 NS 047537-06A1). For this work, MoBa 1 and 2 were supported by the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES-49019) and the Norwegian Research Council/BIOBANK (grant no 221097). We are grateful to all the participating families in Norway who take part in this on-going cohort study.

Availability of data and materials

Access to individual-level Illumina HumanMethyl450 Beadchip data for the MoBa study dataset is available by application to the Norwegian Institute of Public Health using a form available on the English language portion of their website at http://www.fhi.no/eway/. Specific questions regarding MoBa data access can be directed to Wenche Nystad: Wenche.Nystad@fhi.no.

Authors' contributions

Project concept and design: SJL, DMR, AM. DMR was primarily responsible for the data analysis with input from BRJ, SKW, MCW, and SJL and supervision from AM. Data collection: BRJ, SHE, RMN, PMU, WN, SJL. DMR drafted the manuscript. All authors read and approved the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not Applicable.

Ethics approval and consent to participate

The MoBa study has been approved by the Regional Committee for Ethics in Medical Research, the Norwegian Data Inspectorate, and the Institutional Review Board of the National Institute of Environmental Health Sciences, North Carolina, and written informed consent was provided by all participants.

Author details

[1]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [2]Department of Statistics, North Carolina State University, Raleigh, NC, USA. [3]Division of Intramural Research, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, PO Box 12233, MD A3-05, Research Triangle Park, NC 27709,

USA. [4]Norwegian Institute of Public Health, Oslo, Norway. [5]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [6]Centre for Clinical Research, Haukeland University Hospital, Bergen, Norway. [7]Department of Clinical Science, University of Bergen, Bergen, Norway. [8]Laboratory of Clinical Biochemistry, Haukeland University Hospital, Bergen, Norway. [9]Center for Comparative Medicine and Translational Research, North Carolina State University, Raleigh, NC, USA.

Received: 9 February 2016 Accepted: 17 November 2016
Published online: 25 November 2016

References

1. Health UD of, Services H, et al. The health consequences of involuntary exposure to tobacco smoke: a report of the Surgeon General. Atlanta: US Department of Health and Human Services, Centers for Disease Control and Prevention. Coord. Cent. Health Promot. Natl. Cent. Chronic Dis. Prev. Health Promot. Off. Smok. Health; 2006. p. 1988–2002.
2. Bhattacharya S, Beasley M, Pang D, Macfarlane GJ. Maternal and perinatal risk factors for childhood cancer: record linkage study. BMJ Open. 2014;4: e003656.
3. Joubert BR, Haberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450 K Epigenome-Wide Scan Identifies Differential DNA Methylation in Newborns Related to Maternal Smoking during Pregnancy. Environ Health Perspect. 2012;120:1425–31.
4. Lee KWK, Richmond R, Hu P, French L, Shin J, Bourdon C, et al. Prenatal exposure to maternal cigarette smoking and DNA methylation: epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environ Health Perspect. 2015;123:193–9.
5. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet. 2015;24:2201–17.
6. Markunas CA, Xu Z, Harlid S, Wade PA, Lie RT, Taylor JA, et al. Identification of DNA methylation changes in newborns related to maternal smoking during pregnancy. Environ Health Perspect. 2014;122:1147–53.
7. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, et al. Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol. 2014;15:R73.
8. McCarthy MI, Hirschhorn JN. Genome-wide association studies: potential next steps on a genetic journey. Hum Mol Genet. 2008;17:R156–65.
9. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014;15:335–46.
10. Li M-X, Gui H-S, Kwan JSH, Sham PC. GATES: A Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure. Am J Hum Genet. 2011;88:283–93.
11. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. PLoS Comput Biol. 2012;8:e1002375.
12. Lee S, Miropolsky L, Wu M. SKAT: SNP-set (Sequence) Kernel Association Test. 2014.
13. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011;89:82–93.
14. DiFranza JR, Aligne CA, Weitzman M. Prenatal and Postnatal Environmental Tobacco Smoke Exposure and Children's Health. Pediatrics. 2004;113:1007–15.
15. Pang D, McNally R, Birch JM. Parental smoking and childhood cancer: results from the United Kingdom Childhood Cancer Study. Br J Cancer. 2003;88: 373–81.
16. Stjernfeldt M, Lindsten J, Berglund K, Ludvigsson J. Maternal Smoking During Pregnancy and Risk of Childhood Cancer. Lancet. 1986;327:1350–2.
17. Jaakkola JJ, Gissler M. Maternal smoking in pregnancy as a determinant of rheumatoid arthritis and other inflammatory polyarthropathies during the first 7 years of life. Int J Epidemiol. 2005;34:664–71.
18. Aujla SJ, Kolls JK. IL-22: A critical mediator in mucosal host defense. J Mol Med. 2009;87:451–4.
19. Hinks A, Ke X, Barton A, Eyre S, Bowes J, Worthington J, et al. Association of the IL2RA/CD25 gene with juvenile idiopathic arthritis. Arthritis Rheum. 2009;60:251–7.
20. Sakaguchi S, Sakaguchi N, Asano M, Itoh M, Toda M. Pillars Article: Immunologic Self-Tolerance Maintained by Activated T Cells Expressing IL-2 Receptor α-Chains (CD25). Breakdown of a Single Mechanism of Self-Tolerance Causes Various Autoimmune Diseases. J. Immunol. 1995. 155: 1151–1164. J Immunol. 2011;186:3808–21.
21. Stick S, Burton P, Gurrin L, Sly P, LeSouëf P. Effects of maternal smoking during pregnancy and a family history of asthma on respiratory function in newborn infants. Lancet. 1996;348:1060–4.
22. John EM, Savitz DA, Sandler DP. Prenatal Exposure to Parents' Smoking and Childhood Cancer. Am J Epidemiol. 1991;133:123–32.
23. Sasco AJ, Vainio H. From in utero and childhood exposure to parental smoking to childhood cancer: a possible link and the need for action. Hum Exp Toxicol. 1999;18:192–201.
24. Joubert BR, Håberg SE, Bell DA, Nilsen RM, Vollset SE, Midttun Ø, et al. Maternal Smoking and DNA Methylation in Newborns: In Utero Effect or Epigenetic Inheritance? Cancer Epidemiol Biomarkers Prev. 2014;23:1007–17.
25. Porter LA, Dellinger RW, Tynan JA, Barnes EA, Kong M, Lenormand J-L, et al. Human Speedy a novel cell cycle regulator that enhances proliferation through activation of Cdk2. J Cell Biol. 2002;157:357–66.
26. Willett CG, Boucher Y, di Tomaso E, Duda DG, Munn LL, Tong RT, et al. Direct evidence that the VEGF-specific antibody bevacizumab has antivascular effects in human rectal cancer. Nat Med. 2004;10:145–7.
27. Ferrara N, Hillan KJ, Gerber H-P, Novotny W. Discovery and development of bevacizumab, an anti-VEGF antibody for treating cancer. Nat Rev Drug Discov. 2004;3:391–400.
28. Claesson-Welsh L, Welsh M. VEGFA and tumour angiogenesis. J Intern Med. 2013;273:114–27.
29. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000;100:57–70.
30. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
31. Yue W, Yager JD, Wang J-P, Jupe ER, Santen RJ. Estrogen receptor-dependent and independent mechanisms of breast cancer carcinogenesis. Steroids. 2013;78:161–70.
32. Shang Y. Molecular mechanisms of oestrogen and SERMs in endometrial carcinogenesis. Nat Rev Cancer. 2006;6:360–8.
33. Pearce ST, Jordan VC. The biological role of estrogen receptors α and β in cancer. Crit Rev Oncol Hematol. 2004;50:3–22.
34. Miller LL, Pembrey M, Davey Smith G, Northstone K, Golding J. Is the Growth of the Fetus of a Non-Smoking Mother Influenced by the Smoking of Either Grandmother while Pregnant? PLoS ONE. 2014;9:e86781.
35. Aagaard-Tillery KM, Porter TF, Lane RH, Varner MW, Lacoursiere DY. In utero tobacco exposure is associated with modified effects of maternal factors on fetal growth. Am J Obstet Gynecol. 2008;198:66.e1–6.
36. Dai H, Leeder JS, Cui Y. A modified generalized Fisher method for combining probabilities from dependent tests. Front. Genet. [Internet]. 2014;5. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3929847/. cited 1 Oct 2014.
37. Magnus P, Irgens LM, Haug K, Nystad W, Skjaerven R, Stoltenberg C, et al. Cohort profile: the Norwegian mother and child cohort study (MoBa). Int J Epidemiol. 2006;35:1146–50.
38. Rønningen KS, Paltiel L, Meltzer HM, Nordhagen R, Lie KK, Hovengen R, et al. The biobank of the Norwegian Mother and Child Cohort Study: a resource for the next 100 years. Eur J Epidemiol. 2006;21:619–25.
39. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol. 2016;45:382–8.
40. Midttun Ø, Hustad S, Ueland PM. 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. 2009;23:1371–9.
41. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98:288–95.
42. Sandoval J, Heyn H, Moran S, Serra-Musach J, Pujana MA, Bibikova M, et al. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. Epigenetics. 2011;6:692–702.
43. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.
44. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/. [Internet]. Vienna: R Foundation for Statistical Computing; 2014. Available from: http://www.R-project.org.
45. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe

design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29:189–96.
46. Reese SE, Archer KJ, Therneau TM, Atkinson EJ, Vachon CM, de Andrade M, et al. A new statistic for identifying batch effects in high-throughput genomic data that uses guided principal components analysis. Bioinformatics. 2013. doi: 10.1093/bioinformatics/btt480.
47. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
48. JT L, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Storey JD. sva: Surrogate Variable Analysis. R package version 3.12.0. 2014.
49. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100:9440–5.
50. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics. 2012;13:86.
51. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–50.
52. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004;20:289–90.