Early-life factors and epigenetic precursors of childhood leukemia

Alexei Novoloaca

N° d'ordre NNT : 2020LYSE1105

THESE de DOCTORAT DE L'UNIVERSITE DE LYON
opérée au sein de
l'Université Claude Bernard Lyon 1

Ecole Doctorale ED 340
BIOLOGIE MOLÉCULAIRE INTÉGRATIVE ET CELLULAIRE (BMIC)

Spécialité de doctorat : biostatistiques

Soutenue publiquement le 22/06/2020, par :
Alexei NOVOLOACA

Early-Life Factors and Epigenetic Precursors of Childhood Leukemia

---

Devant le jury composé de :

Dr SEVERI Gianluca, Directeur de Recherche, Inserm UMR 1018, Institut Gustave Roussy, Université Paris-Saclay - Rapporteur
Dr SIROUX Valérie, Chargée de recherche, Inserm UMR 5309, Université Grenoble Alpes - Rapporteure
Pr BAROUKI Robert, Professeur des Universités, Inserm UMR-S 1124, Université Paris Descartes - Examinateur
Pr FERVERS Béatrice, Professeure associée, Centre Léon Bérard, Inserm U1052, CRCL, Université de Lyon - Présidente

Dr HERCEG Zdenko, Chercheur/Chef de Section, CIRC, Groupe épigénétique - Directeur de thèse
Dr GHANTOUS Akram, Chercheur, CIRC, Groupe épigénétique - Co-directeur de thèse

Laboratory

Epigenetics Group (EG)
Section of Mechanisms of carcinogenesis (MCA)
International Agency for Research on Cancer (IARC)
World Health Organization (WHO)
150 cours Albert Thomas
69372 Lyon CEDEX 08
FRANCE

Résumé

Le cancer infantile (CI) est la première cause de décès par maladie chez l'enfant, la leucémie (LI) étant le type le plus fréquent. Il s'agit d'une maladie rare dont l'étiologie est peu connue. L'oncogenèse pourrait débuter *in utero*, une période sensible où l'embryogenèse est contrôlée par des régulations épigénétiques héréditaires. Nous avons émis l'hypothèse que des dérégulations épigénétiques (méthylation de l'ADN) *in utero* constituent des mécanismes biologiques sous-jacents associant les facteurs précoces à la LI. Nous nous sommes focalisés sur trois facteurs intrinsèques étroitement liés : le poids à la naissance (l'un des phénotypes les plus précoces prédisposant à la LI), l'âge gestationnel et le sexe de l'enfant.

Nous avons conduit des analyses épigénomiques en utilisant des méthodes biostatistiques optimisées dans des cohortes de naissance et avons trouvé de profondes associations entre le méthylome du sang de cordon et chacun des trois facteurs précoces. Nous avons ensuite étudié si les marqueurs identifiés sont liés significativement à un risque de CI. Enfin, nous avons recherché la proportion du méthylome agissant comme médiateur entre le poids à la naissance et la leucémie. Ces trois étapes constituent une modélisation triangulaire qui vise à identifier des mécanismes moléculaires liant une exposition précoce au cancer.

Ce travail a permis d'identifier des marqueurs épigénétiques de facteurs précoces et de mieux comprendre les dérégulations épigénétiques *in utero* qui pourraient être à l'origine de la LI. Cette approche triangulaire pourra être utilisée pour d'autres études s'intéressant aux effets d'une exposition et aux mécanismes moléculaires sous-jacents.

Abstract

Childhood cancer (CC) is the leading cause of disease-related mortality in children, with childhood leukemia (CL) being the predominant type. CC is rare, and its risk factors and molecular precursors are poorly understood and may originate *in utero*. Fetal life represents a sensitive period during which epigenetic regulation constitutes heritable mechanisms driving embryogenesis. We hypothesize that epigenetic (DNA methylation) deregulation *in utero* underlies biological pathways linking early-life factors to CL. We focus on birthweight, as a collective proxy for early-life exposure and one of the earliest phenotypes predisposing to CL, as well as its closely related intrinsic factors, gestational age and child sex.

We first performed epigenome-wide analysis with optimized biostatistics methodology in large population-based cohorts and found profound associations between each of the three factors and cord blood DNA methylome. Second, we investigated, in a subset of cohorts enriched in CL cases, whether the identified birthweight biomarkers significantly associate with cancer risk and are affected by gestational age and child sex. Third, we tested the proportion by which DNA methylation mediates the effect between birthweight and CL. These steps constitute a proposed 'three-way modelling' aiming to identify molecular mechanisms linking early-life exposure to CC risk.

This work identified epigenetics markers of early-life factors based on some of the largest studies to date, yielding insights into epigenetic deregulation *in utero* that could be at the origin of CL. The described framework could be useful for other exposure-outcome studies investigating underlying molecular mechanisms.

Acknowledgements

First of all, I would like to thank Dr Zdenko Herceg and Dr Akram Ghantous who supervised me during my PhD at IARC. I am very grateful for having given me the opportunity to work with skilled scientists in such an important and motivating field of research.

I could not thank you enough Akram for the devoted supervision during my PhD. I have learned a lot thanks to you, which permitted me to develop both personally and professionally. Your hard work and interest in science is contagious, and you transmit to all of your students your passion for research, your motivation as well as your rigor. You have always been patient and considerate and you advise me many times so that I can improve my skills, enhance my knowledge and provide the best work possible. It is a great honor to work with you.

I am also so grateful to Zdenko who trusted in my potential when I was still completing my Master 2 elsewhere and allowed me to join the Epigenetics Group. Thank you for your support, your pertinent advice and your openness. It allowed me to be involved in interesting projects and international collaborations, yielding fruitful outcomes.

A conducive and friendly working environment is very important for the productivity and well-being. I would like to thank my colleague, Mr. Vincent Cahais, with whom I worked many times during my PhD. Exchanges with you are insightful and very collegial. I am very appreciative of the laboratory work done by Mr. Cyrille Cuenin in order to generate a substantial amount of the epigenome-wide data that was analysed in this thesis. I am also grateful to Dr Claire Renard and Mr Liacine Bouaoun for the enriching discussions we had about biostatistics/bioinformatics and your readiness to help whenever needed. I could also count on the support of Dr Vivian Viallon and Dr Jérémie Becker, my supervisors during my Master 1 and 2 internships. Thank you for teaching me that much about biostatistics and for your availability. Thank you to Elisabeth for having always assisted me many times and very promptly on various administrative procedures. I had the chance to be surrounded by trusted, encouraging and helping people in the Epigenetics Group; thank you each and every one for having contributed in different ways. I would also like to acknowledge the participating children and their families as well as research colleagues from the I4C, CLIC, PACE and EXPOsOMICS consortia.

I am also thankful to Dr Johanna Lepeule and Dr Gaël Yvert who accepted to be members of my thesis follow-up committee. Your availability and willingness to meet with me in person on a yearly basis was for me a very important source of tangible support, advice and guide during my PhD. I am also grateful for the members of my thesis jury for their expertise and precious time; it is an honor that you accepted to evaluate my work and be actively available for my thesis defense.

Finally, I am grateful to have a close-knit family. Immense thanks to my wife and children, to my parents and to my in-laws. A PhD requires personal involvement and wouldn’t have been possible without your love and continued support.

Résumé en français

Le cancer est la première cause de décès dû à une maladie chez les enfants, la leucémie étant le cancer pédiatrique le plus fréquent. L'étiologie du cancer chez l'enfant est encore peu connue. Cela est notamment lié au fait que cette maladie est rare et que les études prospectives sont manquantes. Or, les études prospectives ont une puissance plus forte que les études rétrospectives, potentiellement sources de biais de rappel et de sélection. Une collaboration internationale via des consortia est donc cruciale pour la collecte de données prospectives et d'échantillons biologiques, le Centre International de Recherche sur le Cancer jouant un rôle important dans les consortia majeurs que sont I4C (International Childhood Cancer Cohort Consortium) et CLIC (Childhood Leukemia International Consortium). Les consortia permettent également d'avoir un nombre suffisant d'individus à la fois exposés et malades. Une meilleure compréhension de l'étiologie du cancer chez l'enfant est essentielle afin d'améliorer le diagnostic précoce, mettre au point des thérapies ciblées et prévenir au mieux cette maladie.

Une exposition à certains facteurs extrinsèques a été associée à un risque augmenté de cancer chez l'enfant aussi bien grâce à des études rétrospectives que prospectives. C'est le cas des radiations ionisantes, classées comme carcinogènes, ou encore des pesticides dont le risque concerne aussi bien une exposition professionnelle des parents avant la conception ou pendant la grossesse qu'une exposition résidentielle. Une exposition à des maladies infectieuses pendant la grossesse à quant à elle été associée à un risque moindre de cancer dans les premières années de vie, via un intermédiaire qu'est l'ordre de naissance au sein d'une fratrie. Un facteur intrinsèque, le poids à la naissance, a aussi été utilisé comme intermédiaire entre les expositions auxquelles l'enfant est exposé au cours de la grossesse et sa santé dans les premières années de vie. Un poids de naissance élevé a été associé à un risque accru de cancer chez l'enfant. Il s'agit de l'un des phénotypes les plus précoces de prédisposition à la leucémie infantile.

Concernant les mécanismes biologiques, des analyses systématiques récentes ont montré que peu de mutations (voire aucune mutation) ne sont rencontrées lors de cancer chez l'enfant, dont les leucémies, par rapport aux cancers chez les adultes. Ces résultats mettent en évidence la contribution potentielle des facteurs non génétiques (notamment épigénétiques) au développement du cancer chez l'enfant. L'apparition dès le plus jeune âge du cancer (avant l'âge de 5 ans pour la leucémie) suggère que la maladie pourrait prendre son origine dès la grossesse. Notre attention s'est portée sur la vie intra-utérine, période sensible durant laquelle l'embryogenèse est contrôlée par des régulations épigénétiques héréditaires (dont la méthylation de l'ADN, un des mécanismes épigénétiques majeurs). L'épigénétique étudie les mécanismes impliqués dans le contrôle de l'expression des gènes sans changement de la séquence de l'ADN. Des dérégulations épigénétiques au cours de la grossesse pourraient constituer les mécanismes biologiques liant les facteurs précoces au cancer chez l'enfant.

Des marqueurs épigénétiques variés, associés à différents types de facteurs ou d'expositions ont été identifiés, notamment pour le tabagisme, l'âge, l'âge gestationnel, etc. Notre objectif est d'identifier des marqueurs épigénétiques des facteurs précoces et les mécanismes biologiques sous-jacents permettant de relier les facteurs précoces au cancer chez l'enfant. Nous avons appliqué une approche triangulaire afin d'étudier le lien entre l'exposition à des facteurs précoces, le cancer chez l'enfant et les mécanismes épigénétiques.

Pour ce projet de thèse, nous avons utilisé des échantillons biologiques collectés chez des nouveau-nés sources de données épigénomiques et autres omiques ainsi que des questionnaires d'exposition pendant la grossesse de haute qualité obtenus prospectivement à partir de quatre consortia de cohortes de naissance internationaux. Les analyses ont été réalisées grâce à des technologies de pointe à haut débit et des méthodes bioinformatiques et biomathématiques avancées, afin de tester si une exposition fœtale et des précurseurs moléculaires *in utero* sont des processus responsables plus tard de la leucémie chez l'enfant. Les méthodes utilisées pour le traitement des données sont essentielles et doivent être optimisées afin d'obtenir ensuite des résultats pertinents. Nous avons ainsi participé à trois études qui ont conduit à la publication de deux articles, un troisième étant en révision. L'objectif de la première étude était d'identifier et de corriger les biais présents dans les données de méthylation à grande échelle tels que les effets de lot, qui sont parfois négligés mais qui deviennent problématiques surtout quand ils sont corrélés avec des variables biologiques d'intérêt. Le deuxième article s'est intéressé aux compétences nécessaires pour extraire les informations épigénomiques et des autres types d'omiques, les avancées technologiques récentes facilitant la collection de données omiques à grande échelle à partir de mêmes échantillons biologiques. Enfin, le dernier article représente une investigation pan-cancer utilisant les données génomiques et transcriptomiques pour l'évaluation du caractère moteur d'un ensemble de gènes épigénétiques (analyse intégrative des deux omiques).

Afin de comprendre les mécanismes impliqués dans le développement du cancer chez l'enfant, nous avons recherché les marqueurs épigénétiques associés à trois facteurs intrinsèques étroitement liés : le poids à la naissance, l'âge gestationnel et le sexe de l'enfant. Le poids à la naissance est positivement corrélé à la durée de la grossesse et les garçons ont un poids de naissance plus grand en moyenne que les filles. Enfin, nous nous sommes concentrés sur la leucémie aiguë lymphoblastique, qui est le sous-type le plus fréquent, afin de limiter l'hétérogénéité des données et augmenter la puissance des études. Cette partie de la thèse a contribué à la publication de deux autres articles. Les marqueurs CpGs des facteurs précoces ont été recherchés dans des études basées sur une large population (nombreux sujets non-malades) utilisant une couverture large de l'épigénome (EWAS). Des méta-analyses des résultats EWAS ont ensuite été réalisées pour le poids à la naissance, l'âge gestationnel et le sexe de l'enfant respectivement à partir de 8825, 6885 et 8314 nouveau-nés provenant de 24, 20 et 16 cohortes de naissance au sein du consortium PACE (Pregnancy And Childhood Epigenetics). Nous avons trouvé une association profonde entre la méthylation de l'ADN à la naissance et chacun des trois facteurs intrinsèques. La méthylation de l'ADN dans le sang néonatal a été associée après une correction Bonferroni au poids à la naissance à 914 sites CpGs, à l'âge

gestationnel à 8,899 sites CpGs et au sexe de l'enfant à 46,554 sites CpGs pour les autosomes et 9,372 sites CpGs sur le chromosome X. Une proportion substantielle des signaux disparait au cours de l'enfance et l'adolescence pour le poids à la naissance et l'âge gestationnel, contrairement aux marqueurs CpGs du sexe dont la majorité persiste au cours de l'enfance. Cependant, la persistance de ces marqueurs épigénétiques n'est pas impérative, les évènements épigénétiques qui surviennent à une période critique du développement pouvant avoir des conséquences sur le long terme. Les CpGs de l'âge gestationnel identifiés dans le sang de cordon ont également été retrouvés dans les poumons et le cerveau (ces tissus provenant de feuillets embryonnaires différents). Les marqueurs épigénétiques identifiés dans le sang de cordon permettent donc de capter la plasticité épigénomique du développement prénatal à travers les tissus. Les marqueurs épigénétiques du poids à la naissance ont été étudiés plus en détail dans une nouvelle étude intégrant quatre types de données omiques et ont révélé un rôle important pour la biosynthèse du cholestérol.

Une fois que des marqueurs ont été identifiés, l'objectif est de déterminer si ces derniers sont associés au cancer chez l'enfant. Nous avons ainsi étudié si les marqueurs du poids à la naissance identifiés dans notre étude sont associés à la leucémie chez l'enfant après ajustement des covariables : âge gestationnel, sexe, tabagisme maternel et composition en globules blancs. Parmi les 8696 sites CpGs du poids à la naissance, 414 sont associés à la leucémie chez l'enfant ($p<0,005$). De plus, deux analyses de médiation différentes ont indiqué un CpG sur un gène non codant qui associe le poids à la naissance et la leucémie (p ajusté = 0.0037) et qui n'est pas confondu par l'âge gestationnel ou le sexe de l'enfant. Le CpG identifié ne faisait pas partie des marqueurs épigénétiques de l'âge gestationnel. La vérification de ces résultats nécessite la réplication de l'analyse dans d'autres cohortes et une vérification expérimentale grâce à des essais fonctionnels.

Nos résultats ont permis d'identifier de potentiels mécanismes biologiques reliant le poids à la naissance et ses facteurs intrinsèques étroitement liés (durée de grossesse et sexe de l'enfant) à la leucémie infantile. La réplication de ces résultats et une analyse fonctionnelle approfondie par le biais de modèles expérimentaux pourront aider à déterminer le rôle des marqueurs épigénétiques identifiés et à caractériser les gènes moteurs et les voies causales impliqués dans l'oncogénèse du cancer infantile. Notre approche triangulaire pourra être utilisée pour d'autres études s'intéressant à des effets de l'exposition et aux mécanismes moléculaires sous-jacents.

List of acronyms

ALL	Acute Lymphoblastic Leukemia
AML	Acute Myeloid Leukemia
BMI	Body Mass Index
BW	Birthweight
CC	Childhood Cancer
CCLS	California Childhood Leukemia Study
CL	Childhood Leukemia
CLIC	Childhood Leukemia International Consortium
CNS	Central Nervous System
CpG	Cytosine phosphate Guanine
CT	Computed Tomography
DNA	Deoxyribonucleic Acid
DOHaD	Developmental Origins of Health and Disease
DMR	Differentially Methylated Region
DNMT	DNA MethylTransferase
EPIC	European Prospective Investigation into Cancer and Nutrition
ERG	Epigenetic Regulator Gene
EWAS	Epigenome-wide association studies
FDR	False Discovery Rate
GWAS	Genome-wide association studies
HAC	Histone ACetyltransferase
HDAC	Histone DeACetylase
HIC	High Income Countries
HIMA	High-dimensional Mediation Analysis
IARC	International Agency for Research on Cancer
I4C	International Childhood Cancer Cohort Consortium
IR	Ionizing Radiation
MDS	Multi-Dimensional Scaling
MoBa	Mother, Father and Child Cohort Study
MR	Mendelian randomization
LMIC	Low- and Middle-Income Countries
ORA	Overrepresentation Analysis
PACE	Pregnancy And Childhood Epigenetics
PGC	Primordial Germ Cells
PM	Particulate Matter
QC	Quality Control
RE	Repetitive Elements
RIGI	Radiation Induced Genomic Instability
RNA	Ribonucleic Acid
SNP	Single-Nucleotide Polymorphism
SVA	Surrogate Variable Analysis
TET	Ten-Eleven Translocation
TCGA	The Cancer Genome Atlas
WBC	White Blood Cell
WCRF	World Cancer Research Fund
WHO	World Health Organization

Publication list and contribution

1. Sexton-Oates A, **Novoloaca A**, Ghantous A[#], Herceg Z[#]. Cancer. *Environmental Epigenetics in Toxicology and Public Health*. Submitted.
2. Perrier F, **Novoloaca A**, …, Ghantous A, …, Herceg Z, et al. Identifying and correcting epigenetics measurements for systematic sources of variation. Clin Epigenetics. 2018 Mar 21;10:38.
3. Chauvel C*, **Novoloaca A***, et al. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 2019 Feb 14. pii: bbz015.
4. Halaburkova A, Cahais V, **Novoloaca A**, Khoueiry R, Ghantous A, Herceg Z. Pan-cancer genome and transcriptome analysis and orthogonal experimental assessment of epigenetic driver genes. Accepted in Genome Research.
5. Küpers LK, …, Ghantous A, …, **Novoloaca A**, …, Herceg Z, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 2019 Apr 23;10(1):1893.
6. Merid SK*, **Novoloaca A***, Sharp GC*, Küpers LK*, Kho AT*, …, Herceg Z, …, Ghantous A, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. Genome Med. 2020 Mar 2;12(1):25.
7. Solomon O, …, **Novoloaca A**, …, Herceg Z, …, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in newborns and children show widespread sex differences in blood DNA methylation Child sex and DNA methylation. In preparation.
8. Alfano R, Chadeau-Hyam M, Ghantous A, …, **Novoloaca A**, et al. A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism. Metabolism. 2020 Sep;110:154292.
9. Methylation mediates the known association between BW and Childhood Leukemia. **Novoloaca et** al. In preparation.

[#]Equal contribution.
*Co-first authorship.

Papers	1	2	3	4	5	6	7	8	9
data preparation			✓	✓	✓	✓	✓		✓
defining the analysis strategy		✓	✓	✓	✓	✓	✓	✓	✓
statistical analysis			✓	✓	✓	✓	✓		✓
figure generation			✓	✓					✓
manuscript design			✓						✓
writing	✓		✓						✓
proofreading the manuscript	✓	✓	✓	✓	✓	✓	✓	✓	✓

Review
Methodology
Results finished
Results ongoing

Table of Contents

I. Introduction

A. Childhood cancer

Childhood cancer (CC) refers to malignancies that occur in children between birth and the age of 15 to 20 years (variation of the cut-off among countries). The incidence of CC in the world in 2015 was estimated at 360000 cases, mainly in Asia and in Africa (respectively 54% and 28%, results shown in **Figure 1**) (1). A large European multicentric database founded by the International Agency for Research on Cancer (IARC) (2) shows that in recent years rates have increased in Europe by 1-2%/year for most CC types, even though the causes of this increase are not understood (3).

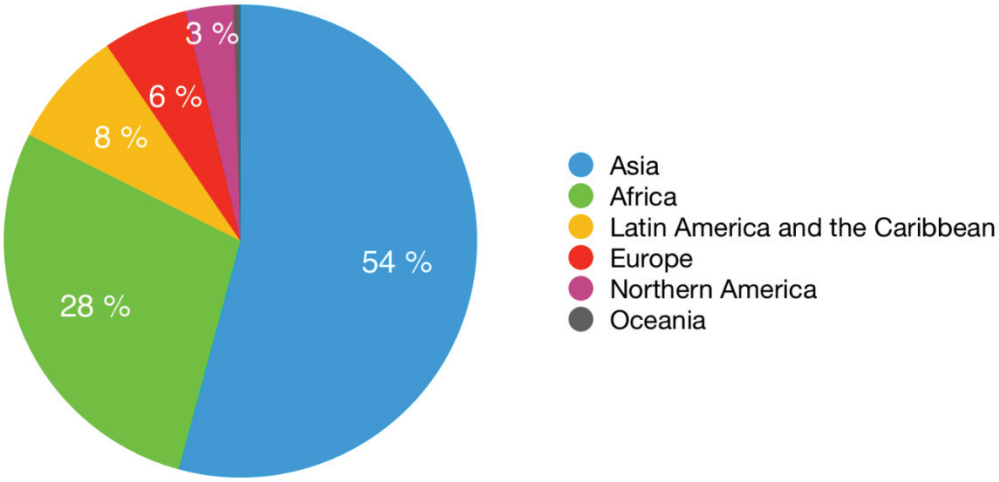


***Figure 1**. Estimated incidence of CC by continent in children aged <15 years in 2015. Values obtained from (1).*

CC is the second cause of death in children after accidents. Although cancer death rates have declined in the last years, cancer remains the disease that causes the most death among children, and a child dies of cancer every three minutes worldwide. The reasons for this are manifold.

First, CC is a very heterogeneous group of malignancies. According to the American Cancer Society, the most common cancer types in children (0-14 years old) are childhood leukemias (CL, 29%), brain and other nervous system tumors (26%) followed by lymphomas and reticuloendothelial neoplasms (12%). The distribution is quite different in adolescents (aged of 15 to 19 years) as the most frequent cancers are brain and other nervous system tumors (21%), lymphomas (20%) and leukemias (13%) (4). This complexity is further compounded when considering the histological subtypes (main subtypes of CL: acute lymphoblastic leukemia (ALL, 75%) and acute myeloid leukemia (AML,19%))(4) as well as specific cytogenetic groups characterized

by chromosomal changes (translocation, deletion, and hyperdiploidy). While some CC are relatively easy to recognize (e.g. Burkitt's lymphoma), other subtypes tend to present non-specific symptoms like fever (e.g. leukemia) or vomiting and weight loss (e.g. brain cancer) which adds another difficulty for the diagnosis of CC cases. Moreover, CC cannot be always treated like adult cancers. The subtypes of CC are distinct from those in adult cancers or may be not occur at all in adult ages. Additionally, cancer and its treatments may have different effects on children compared to adults and the growing bodies may respond differently to drugs.

Second, CC is a rare disease, as reported by IARC (5), and adequately powered prospective studies are lacking which would be the preferred design. Most of the causal evidence comes from retrospective case-control studies, which incur two major limitations: recall bias and high likelihood of control selection bias.

Lastly, there are differences between high income countries (HIC) and low- and middle-income countries (LMIC) in cancer incidence, diagnosis, treatment and mortality partly because LMIC health systems are insufficiently equipped. Approximately 8 in 10 of CC cases live in LMIC, and their survival rate is often near 20%. This is in sharp contrast to HIC, where cure rates exceed 80% for many common CC (6,7). In 2018, the World Health Organization (WHO) established the WHO Global Childhood Cancer Initiative aiming to achieve at least 60% survival for all children diagnosed with cancer around the world by 2030, although this initiative will not address potential environmental causes underlying the disparities in the incidence of CC worldwide.

The number of survivors from CC continues to increase. According to Erdmann et al. five-year survival rates from CC have increased from 30% in the 1960s to 80% nowadays in HIC, with striking differences in survival between cancer types. However, even if the improvement in survival resulted in a decrease in mortality rates, it also led to diversification of adverse effects in survivors during the life-course including increased risk of developing a second malignancy. Hence, there is an urgent need in better understanding of the etiology of this disease that will help enhancing early diagnosis, targeted therapy and prevention.

The rarity of this disease, the various heterogeneous subtypes and the non-specificity of the symptoms are challenges faced by CC epidemiology in order to identify potential underlying risk factors of pediatric cancer. As most CC occur at early age (e.g. before 5 years for leukemias), it suggests that some cancers may start in the womb. Prenatal, fetal and early childhood exposures to extrinsic (ionizing radiation, pesticides and birth order) and intrinsic (birthweight) factors are potential CC risk factors. Emphasis is given to these particular risk factors because they are proved in both retrospective and prospective child cancer consortia, and we will describe each of them in the following sections.

B. Risk factors during early life

1. Ionizing radiation

Ionizing radiation (IR) occurs in two forms, waves (X-rays and gamma rays) and particles (alpha, beta, proton and neutron), and can be natural (80%; cosmic rays, solar rays, radon gas) or artificial (20%; primarily medical imaging like computed tomography (CT) scans, cancer radiotherapy and also nuclear power generation or even from military purposes) (9). Exposure to IR varies due to geographic factors but, as background IR is ubiquitous, zero exposure does not exist. The international units of measure for absorbed dose and equivalent doses are respectively the gray (Gy) and the sievert (Sv). Doses are rarely above 100mGy during diagnostic X-rays and CT scans, but they may exceed 50Gy during radiotherapy treatment of CC (10).

IR is primarily genotoxic and the different types of IR were classified as Group-1 carcinogen by the IARC Monograph, although some uncertainty remains about carcinogenicity of low dose levels (<100mSv) (11). However, a recent pooled study of nine cohorts examined the risk of CL with IR from various sources, and detected an increased risk at exposures <50 mSv and a threefold risk for AML and almost six-fold for ALL with each increase of 100 mSv. This is so far the most convincing evidence of a low-dose IR effect at least in CL (12).

The effects of IR were first investigated after the atomic bombings in Japan (13). The carcinogenic effects of IR are stronger in children, as they are more radiosensitive and have more years of potential life to express the risk. After the Hiroshima and Nagasaki atomic bomb detonations, increased rates of CL were identified 5–6 years later. Similarly, for solid cancers during childhood and adult life, studies of cancer incidence in survivors showed an increased risk in early childhood (up to six years), but not during the prenatal period (14). Interestingly, diminished solid cancer survival observed in children with decreasing age at time of exposure pointed to the importance of early-life exposures to lifetime cancer risks (15). Radio-iodine IR from the 1986 Chernobyl nuclear power plant accident resulted in increased rates of childhood thyroid cancer. For the recent Fukushima Daiichi nuclear accident, thyroid doses were manifold lower and an increase in thyroid cancer has so far been attributed to over-diagnosis (16).

IR used for cancer treatment (radiotherapy) has been identified as a risk factor for CC. CC patients treated with radiotherapy have higher incidence of secondary cancers during childhood and young adulthood than both the general population, and those children treated with chemotherapy only (17–19). They exhibit an increased risk for development of secondary central nervous system (CNS), breast, thyroid, bone and soft tissue cancers (20). Strikingly, radiotherapy treatment for childhood Hodgkin's lymphoma in particular, showed an increased lifetime risk of developing breast cancer, higher in these patients than in women carrying the well-known BRCA1 mutation (21).

Prenatal exposure to radiation through diagnostic X-ray imaging was also associated with an increased risk of CC in a case-control study of 1,500 children (22). A detailed review in 2008 found on the contrary very little evidence for the association between

prenatal X-ray exposure and CC (23). These results however, do not necessarily contradict previous evidence as there are limitations in study design, study size, exposure measurement and involve very low exposures. Nevertheless, X-ray exposure should be limited during the pregnancy especially when other imaging techniques such as ultrasound may be used (24). Childhood X-exposure was also demonstrated to be associated with development of CL (25–29). Increasing use of CT scanning in children in most HIC raised awareness on its possible deleterious effects (10), although potential reverse causation is a concern. A dose-dependent association after exposure to CT scanning in childhood was established in several large studies including a population-based study of 10 million children and adolescents indicating an increased risk of many cancers including CL and brain cancer (30,31). While there is no doubt CT is an important diagnostic instrument, results may urge better dose adjustment when planned examinations with children are conducted.

Mechanisms of IR-induced carcinogenesis have been well characterized. Exposure to IR results in a variety of molecular damage in the cell, including single- and double-strand breaks in DNA that can lead to mutations and chromosomal aberrations (32) and subsequently to cell death or oncogenic transformation. IR-induced molecular changes may result directly in the activation of oncogenes or in the silencing of tumor suppressor genes promoting the malignant properties of normal cells. IR, in addition to being capable of producing mutations, can also induce epigenetic changes. Studies on the non-targeted effects of IR in unexposed cells or progeny of exposed cells (32,33) have revealed a role for epigenetic mechanisms such as radiation induced genomic instability (RIGI), which can also lead to the development of cancer despite never being exposed to radiation (34). RIGI produces mutations, chromosomal instability and gene amplification in the progeny of exposed cells, and it is hypothesized that these effects may be the result of inherited epigenetic signatures from the original exposed cells. Studies were conducted in animal and human models revealing alterations in DNA methylation patterns when exposed to IR. Genome-wide hypomethylation, genome instability, promoter specific hypermethylation and repression of tumor suppressor genes were observed in whole body irradiation with X-rays of rats, mice, hamsters and human cell lines (35–37). Furthermore, exposure to plutonium in humans may increase the risk of adenocarcinoma through hypermethylation by inactivation of a key regulator of the cell cycle (*P16* gene) (38).

2. Pesticides

According to WHO, pesticides are « chemical compounds used to kill pests (insects, rodents, fungi and unwanted plants (weeds)) » (39). They have been used worldwide for decades notably in agriculture to reduce crop losses. They are potentially toxic for humans and can have unwanted side-effect leading sometimes to cancer. More than 20 pesticides have been classified as at least “probable or possible carcinogens” by IARC. Studying the risks of pesticides exposures is difficult as it comprises a heterogenous group of agents, with hundreds of pesticides and thousands of formulations (40).

Given that pesticides are widely used, people can be exposed to them from a large variety of sources. Humans can first of all be exposed at home both indoors and outdoors (lawn and garden, pesticides laden dust, pet insecticides, insecticidal shampoos for lice infestation…) or by eating or drinking contaminated food and water. They can also be exposed to pesticides during their professional activity (e.g. people working in farming and manufacturing). It is noteworthy that farming not only exposes farmers and their family but also contaminates surface water, grounds, crops, etc (41).

Children can be exposed to pesticides through different ways. First, directly, as young children often put objects and their hands in their mouths and they spend a lot of time on the floor (42). There is also a possible indirect contamination from parental exposure before conception (from both parents) and during pregnancy. Paternal exposure might lead to germ cell damage while maternal exposure during pregnancy can result in fetal exposure. Consistent with this notion, pesticide residuals have been found in umbilical cord blood and meconium (43).

Different studies have been conducted to assess if pesticides are environmental risk factors of CC and in particular of CL and CNS tumors. Environmental factors are generally difficult to measure accurately, notably in a retrospective setting (different biases are possible, as mentioned earlier) and there are very few prospective studies available. These studies are further hampered by the existence of numerous pesticides and lack of specific information about pesticide exposure. At last, given the low numbers of CC every year, individual studies often lack statistical power (44).

In 2011, a systematic review and a meta-analysis were conducted in order to estimate the risk of residential exposure to pesticides and CL. Residential use of pesticides has been significantly associated with CL, with the greatest risk when the mothers were exposed during pregnancy. The strongest risk was for indoor exposure, for exposure to insecticides and their link to AML. The authors pointed out that the data were too scarce for causality assessment but it is important to take preventive measures and reduce the use of indoor insecticides during pregnancy (45). More recently, in 2015, Bailey and al. pooled data from 12 case-control studies participating in the Childhood Leukemia International Consortium (CLIC) (46) which permitted to have almost 8,000 leukemia cases and 15,000 controls. They found that parent's pesticide exposure in the few months preceding conception and during pregnancy was associated with a higher risk of ALL and AML. Exposure after birth was associated with an increased risk of ALL but not AML. Little variation was noted by type of pesticide. Hence, they recommended that parents limit pesticide exposure in home during the year before birth and in the first years of childhood (47).

In order to investigate CL risk of parental occupational exposure to pesticides in the prenatal period, Bailey and al. pooled data from 13 case-control studies participating in the CLIC. Maternal occupational exposure during pregnancy has been found to significantly increase the risk of AML (almost doubling risk). Paternal occupational exposure close to conception was associated with a slightly increased risk (stronger effect when diagnosis was established after 5 years old and for children suffering of T cell ALL) (48). A meta-analysis was conducted from 20 case-control and cohort

studies to evaluate the risk of parental exposure and the occurrence of brain tumors. Parental occupational exposure to pesticides was associated with an increased risk of brain tumors, especially for those exposed during prenatal period (for both parents). This study supports the recommendation of minimizing parental occupational exposure to pesticides (49). At last, a prospective study was conducted in the International Childhood Cancer Cohort Consortium (I4C) (50) to assess the risk of parental occupational exposures to pesticides for both CL and CNS tumors. It is a large prospective study as the data were collected on almost 330,000 participants from birth cohorts in five countries. The results showed that paternal exposures to pesticides were associated with higher risk of AML. However, exposures to pesticides did not increase the risk of CNS tumors and ALL. The risk of maternal exposure to pesticides could not be evaluated because of low exposure prevalence in pooled cohorts (51). In summary, the reported studies point out to an increased risk of CL associated with parental occupational exposure as well as residential exposure to pesticides of parents before conception and during pregnancy.

Mechanisms of pesticide-induced carcinogenesis potentially include oxidative stress, genotoxicity and/or epigenetic changes (52). An in-vitro study showed that pesticides may modify gene promoter DNA methylation levels and suggests that epigenetic mechanisms may mediate the effect of pesticide exposure on cancer (53). Pesticide exposure in humans could also induce oxidation of guanine leading to DNA damage, based on a study on soybean farmers (54). People affected by occupational exposure also displayed genomic hypermethylation of DNA, which correlated with micronucleus frequency (54). A systematic review highlighted synergistic interactions in a small subset of pesticides mixtures (55).

3. Birthweight

Birthweight (BW) is a marker of prenatal growth which can be used as a proxy of cumulative effects of in utero exposures and of later health outcomes, including cancer. The potential links between birth characteristics (and more precisely BW) and CC were first suggested in the early 60s by MacMahon and Newill (56). Subsequently, this putative association was investigated using retrospective case control studies (57–62).

A pooled (and similarly a meta-analysis) of 12 case-control studies participating in the CLIC demonstrated an increased risk of ALL for children who were large for, relative to appropriate for, gestational age (61). Another meta-analysis focusing on child neuroblastoma in 10 case-control studies and one cohort showed an increased risk for this cancer type with high BW, while the evidence for the association with low BW was less robust (58). Childhood brain tumors, specifically astrocytoma and medulloblastoma but not ependymoma, have also been positively associated with high BW in a meta-analysis of 8 studies encompassing 1,748,964 children, including 4,162 brain tumor cases (59). An updated meta-analysis on the same topic confirmed the association between high BW and astrocytoma and the absence of an association of high or low BW with ependymoma. The association of BW with medulloblastoma and/or primitive neuroectodermal tumors was inconclusive

(with a trend towards increased risk for both high and low BW) (62). Still, all the evidence presented so far is based on retrospective studies, with likelihood of recall and control selection bias. CC is rare, and well-powered prospective studies have been limited. More recently, the largest birth cohort study to date encompassing 112,781 live births (including 377 cancer cases) from six geographically diverse cohorts by the I4C showed that high BW increases the risk of CL and overall cancers; high BW was also positively associated with non-leukaemia cancer among children diagnosed at age ≥ 3 years but not at younger age (63). Other single cohort or population-based registry studies have reported similar associations but are not covered herein due to the rarity of cases in individual studies (64–66). Overall, based on both retrospective and prospective evidence, BW represents one of the earliest predisposing phenotypes of CC.

Additionally, BW was shown to be associated with outcomes occurring much later in life, including adult cancers. According to the World Cancer Research Fund (WCRF), there is “probable increased risk” of premenopausal breast cancer and “limited suggestive increased risk” of malignant melanoma by high BW (67), while the evidence pertinent to other cancer types is existent but not as solid.

All these studies point out the importance of birth characteristics and of the fetal life, directly or indirectly, on the risk of different cancer types. To add a new dimension to the value of the current epidemiological studies, the next step and missing link is to investigate the molecular mechanisms underlying the associations between BW and cancer. Moreover, molecular studies can additionally generate biomarkers of exposure that can enhance exposure assessment in order to reinforce previous associations or identify new ones. Epigenetic mechanisms are particularly important in fetal life as they represent heritable mechanisms driving embryogenesis as well as molecular sensors of the environment (52). Our new study conducted in collaboration with the Pregnancy And Childhood Epigenetics (PACE) (68) Consortium, based on almost 9,000 neonates from 24 birth cohorts worldwide, shows thousands of differentially methylated markers in association with BW, among which 1.3% persist in childhood and adolescence (69).The study adds to the increasing evidence underscoring the importance of epigenetic mechanisms during early life.

4. Birth order

One of the prevailing theories on the etiology of CL outline the "delayed infection hypothesis", first formulated in 1988 (70), which states that reduced exposure to infection in early life yields “untrained” or naïve immune cells which, once affronted later by a microbial infection, would respond by an exaggerated hyper-proliferation and inflammation, eventually leading to cancer. Evidence on CL occurrence and direct exposure to specific infections (e.g. influenza) is lacking due to the difficulty to precisely measure them and due to the recall bias in retrospective studies. To pursue the investigation between infectious diseases and CL, proxies with evidence for their association with infection should be prioritized. Indeed, different proxy measures have been evaluated, including daycare attendance, breastfeeding, vaccination history, hospitalization or prescriptions for infection, as well as birth order

(71–73). In particular, birth order was associated with an increased risk of common infections found in blood (74). Additionally, birth order has also been used as a proxy for different hormonal exposures to the fetus, and higher birth order children might have higher levels of microchimerism (75).

Birth order was first studied in relation to CC in 1962 by MacMahon and Newill who identified a decrease in CC mortality for increasing birth order (56). Then, further studies tried to unravel the relationship between birth order and CC in general or leukemia, generating, however, heterogeneous results (71,76,77) probably due to differences in study design or inconsistency in population characteristics. More recently, it has been shown that being later born (i.e. being more likely exposed to in utero infections due to presence of other siblings at the time of pregnancy) protects against CL, based on the largest prospective epidemiological study so far (78). Moreover, the association between birth order and CL was significantly strengthened by the interaction with BW (<3kg) and fathers age (30 or older). These two factors have been previously labeled as established (for BW (63) and potential (for parental age (79)) risk factors of CL. Long time considered as potential confounders for the association between birth order and CC, they are now assessed as modifiers of this association enhancing the investigation of etiological pathways leading to CL.

Evidence for the birth order and epigenetics association is lacking, but this field is gaining interest as it can potentially represent the underlying mechanisms leading to later onset of diseases including CC. Multiple births and in particular twin birth were previously investigated to determine DNA methylation (one of the major types of epigenetics) in subsequent siblings (80) because of the changes caused to the intrauterine environment as the uterus is enlarged far more than with a singleton pregnancy. The results show different DNA methylation of siblings born before versus after a twin birth implying the possibility of a different disease (e.g. cancer) risks in later life. However, there is no direct evidence that twin birth changes the intrauterine environment. Hence, there is need of further studies to elucidate the mechanisms underlying these observations. One particularly appealing and relevant future aim would be to catalogue robust epigenetic markers of birth order in the same way it was previously done for BW.

C. Epigenetics

1. Major epigenetic mechanisms

Epigenetics was first defined in the early 1940s by Conrad Hal Waddington as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” (81). Since then, the meaning of the word gradually evolved in “heritable changes in gene expression that are not due to any alteration in the DNA sequence” (82) and finally be narrowed to “the study of changes in gene function that are mitotically and/or meiotically heritable and that do not entail a change in DNA sequence” (83). It was rapidly acknowledged that epigenetics plays an important role in the control of gene expression (84) and development of disease, including cancer (85,86). Evidence highlighting the crucial

role of epigenetics accumulated in the recent years due to the development of powerful technologies and optimized biostatistical and bioinformatical methods.

There are three major types of epigenetic regulations encompassing (i) posttranslational modifications of histone proteins and chromatin remodeling, (ii) non-coding RNA interference and (iii) DNA methylation, all of which create an intricate and self-reinforcing interactions converging on a common cellular process i.e. regulating gene expression (although expression-independent effects of epigenetic mechanisms are being increasingly reported) (**Figure 2**) (87–89).

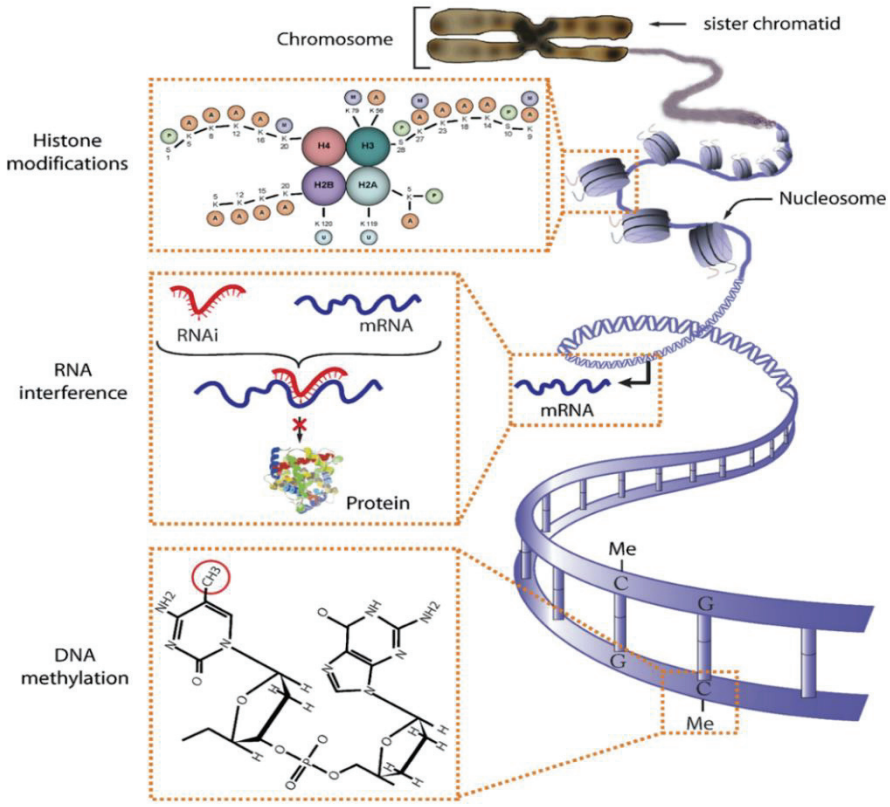


***Figure 2**. Three majors types of epigenetic regulations, adapted from (90)*

Post-translational histone modifications refer to chromatin structure and in particular the histone protein octamer (2 of each histones H3, H4, H2A and H2B around which approximately 1.75 turns of DNA are coiled) that constitute the nucleosome. The nucleosome is hence made of histones and DNA and it is the first level of DNA compaction. Other types of proteins permit additional levels of compaction. Specific enzymes will act in distinct ways producing acetylation, phosphorylation, methylation and ubiquitylination of histones. The modifications will directly affect chromatin structure, or provide dynamic binding platforms for proteins with specific binding domains. Chromatin remodeling concerns the restructuring of nucleosomes within chromatin controlling the access to DNA. Indeed, DNA transcription into RNA is facilitated when chromatin is decondensed as it permits the access to DNA to the transcription complex. It plays an important role in several key biological processes,

including DNA replication and repair, apoptosis, development and pluripotency (91) and it has been associated with human diseases, such as cancer (92,93). Histone acetyltransferase (HAC) and histone deacetylase (HDAC) have notably an impact on DNA transcription. Histone deacetylation is generally linked to transcriptional inhibition (94).

Although non-coding RNA remains the less studied mechanism in comparison with the two others, it also participates in regulating gene function. Most of the time, non-coding RNAs act as molecular guides, and their huge number in typical cells certainly represents a crucial role in regulating all DNA processes. MicroRNAs (miRNAs) have been previously implicated in epigenetic inheritance across generations (95).

DNA methylation is the most extensively studied epigenetic mechanism, with abnormal DNA methylation levels occurring in almost all human cancers (96). DNA methylation refers to the addition of a methyl group to the 5′ position of the cytosine pyrimidine ring. This reaction is mediated by particular enzymes named DNA methyltransferases (DNMTs). There are 3 major DNMTs: DNMT1, a maintenance DNMT, and DNMT3a and DNMT3b which are *de novo* DNMTs. DNMT1 has a role in DNA methylation maintenance through mitosis. After DNA replication, the synthetized strand is not methylated contrary to the parent strand. DNMT1 binds to CpG sites on the synthetized strain and ensures the methylation of cytosines in a pattern matching that of the parent strand (97). The *de novo* DNMT3a and DNMT3b have a key role in de novo methylation of DNA particularly during embryogenesis or later developmental processes that involve tissue differentiation (see section I.C.2.) (98). Methylated cytosine are often localized in regions rich in cytosine and guanine, called CpG islands, but can also occur in non-CpG contexts (99). DNA methylation patterns are faithfully propagated through cellular division and are hence stable and mitotically heritable. Aberrant DNA methylation changes are also stably propagated through cell division although they are, in contrast to mutations, potentially reversible. Removal of methyl group (demethylation) can be both active, with the intervention of TET family enzymes or passive, when DNA is replicated in the absence of maintenance DNA methylation by DNMT1 of newly synthesized DNA strands (100). Regions flanking CpG islands are called shores (< 2kb flanking CpG islands) and shelves (from 2kb to 4 kb of CpG islands). The rest of the genome is called open sea. Methylation is more dynamic along the shores and shelves. It has notably been shown that differences of methylation pattern between tissues or between normal versus tumor cells often occur at shores rather than at CpG islands themselves (101,102). Cancer cells, unlike healthy cells, present a global DNA hypomethylation accompanied by hypermethylation in promoter regions of specific genes, among which the tumor suppressor genes are frequently affected (103).

The mechanisms by which DNA methylation or demethylation lead to transcription silencing or activation are not necessarily identical, depending on the genomic site of occurrence (eg. gene promoters, gene bodies, repetitive elements, etc.). DNA methylation of promoter sequences often down-regulates gene expression; whereas, gene-body methylation has been positively correlated with gene expression. Paradoxically, DNA methylation is more prevalent within gene-bodies compared to promoters (104). Gene-body methylation levels are supposed to be predominantly

shaped via the accessibility of the DNA to methylating enzyme complexes (105). DNA methylation also impacts chromatin compaction, preventing the access to DNA to the transcription complex. In addition to regulating gene expression, DNA methylation is implicated in X-chromosome inactivation, gene imprinting and transposon silencing (106–111). About 50% of the human genome is composed of DNA repetitive elements (RE) which are relics of transposons. They can proliferate and move throughout the genome (112). DNA Methylation in RE hampers their mobility and maintains genomic stability. Decrease in methylation in RE is frequently observed in tumor (113).

2. Epigenetics in the early-life period

Not only the type of environmental exposure, but also its timing plays an important role in influencing disease risk. Fetal life represents an exposure-sensitive period in the human life course during which epigenetic regulation constitutes heritable mechanisms driving embryogenesis. It is hypothesized that epigenetic deregulation *in utero* lies at the heart of causal pathways linking early-life factors and CC (114).Fetal life is hence an important period as changes in cell fate during embryonic development could potentially have lifelong health consequences.

Epigenetics can shed light on the mechanisms for the developmental origins of health and disease (DOHaD) approach (115). This concept aims to decipher biological pathways underlying existing epidemiological evidence linking early life exposure with later onset of diseases or to identify new risk factors by using epigenetic biomarkers as their proxies (116,117). Studying epigenetic mechanisms during development might explain the underlying causes of some diseases (109).

The rapid advancement in computational approaches together with cutting-edge laboratory technologies allowed epigenetics to be used as a predictor for tobacco smoking status (as well as duration) (118), child sex (in preparation) , age (119), gestational age (120), cell type composition (121) and ethnicity (122). Numerous life-course exposures were studied in relation to the epigenome expanding our knowledge of epigenetic signatures.

DOHaD rationale particularly emphasizes the importance of the profound epigenetic reprograming (detailed below) which occur in early embryonic development and germ cell specification, during which methylation landscape is greatly remodelled (**Figure 3**). Defects in this machinery have important repercussions on embryonic development and later life health (123).

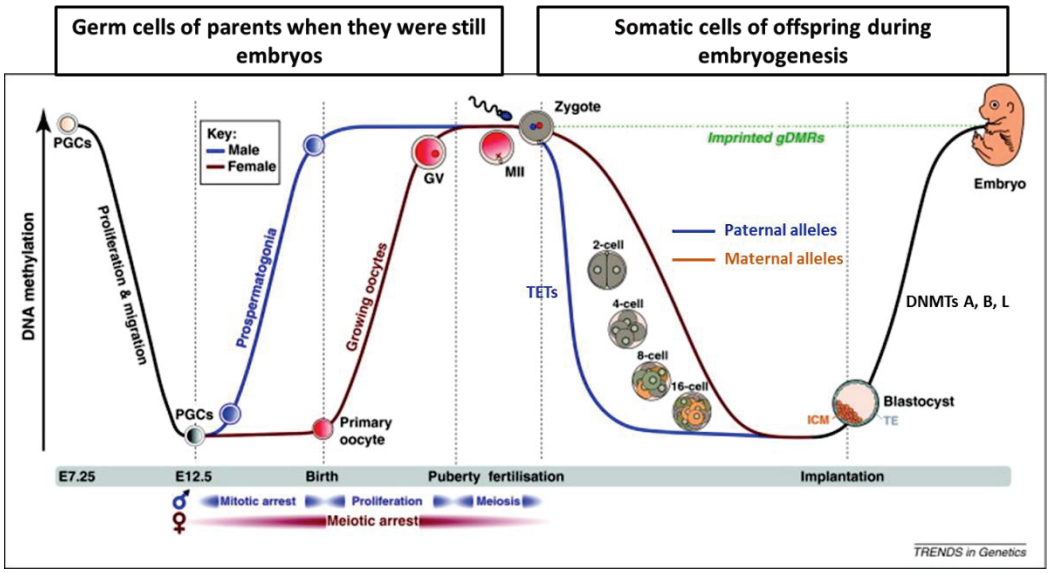


***Figure 3**. Epigenetic reprogramming in early embryonic development and germ cell specification, adapted from (124)*

As shown in **Figure 3**, when the zygote is produced, demethylation of both paternal and maternal alleles initiates and continues in the daughter cells that are generated from the mitotic divisions afterwards to reach the blastocyst stage. The paternal alleles are actively demethylated by the action of TET enzymes that originate from the oocyte while the maternal alleles are demethylated passively by mitotic divisions in the absence of maintenance methylation. The only exceptions that escapes demethylation at this stage are imprinted genes. At blastocyst stage, remethylation initiates and further progresses in subsequent cell generations producing distinct methylome patterns in the different types of tissues, which include the primordial germ cells (PGCs) of the embryo. PGCs then undergo a second wave of global demethylation which also encompasses the imprinted genes, hence, erasing any parent-of-origin pattern of methylation. Once demethylation is complete in the PGCs, remethylation starts again in those cells in a sex-specific manner in order to give rise to gametes that contain sex-specific epigenetic signatures. In males, remethylation is reinitiated *in utero* while the PGCs are mitotically arrested and is completed at birth. Afterwards, PGCs proliferate during the lifetime and undergo meiosis starting at puberty in order to produce sperms. In females, remethylation starts at birth during meiotic arrest and is completed at puberty when meiosis resumes to produce oocytes. Although embryonic and germline demethylation is global, a substantial amount of DNA methylation escapes the erasure process, hence, serving as a potential vector of epigenetic inheritance (124).

3. Epigenetics and childhood cancer

As CC can be diagnosed only few years after birth, the chance of acquiring enough mutations to drive oncogenesis is smaller than in adult cancers. This reinforces the earlier mentioned assumption that initiating events may have occurred around the in-utero period. Recent large-scale genetic sequencing of childhood tumors identified

few or no mutations (125–127). Pediatric cancers harbor a low mutational burden compared to adult cancers (128). These findings highlight the potential contribution of non-genetic (notably epigenetic) factors to CC development (114). Epigenetic mechanisms play an important role in cancer development and progression (129) mediating gene-environment interactions and their effect throughout the tumorigenesis process (130), although the contribution of deregulated epigenetic mechanisms to childhood cancer is not well understood.

Abnormal epigenetic states (induced by altered DNA methylation pattern, mutations in histones, etc.) are important for initiation and progression of many CCs. Factors influencing the oncogenic potential of epigenetic alterations include the particular development stages and specific cellular types in which the alterations occurred (128). The cell of origin in which the oncogenic event happens has a major importance for determining the phenotype of cancer, notably for leukemia (131). A large proportion of various CC types displayed alterations in genes encoding epigenetic regulators and chromatin complexes in a pan-cancer analysis (125,126). Epigenetic regulators refer to the genes whose products change the epigenome directly through DNA methylation, post-translational modification of chromatin or alteration of the structure of chromatin. They are frequently the target of mutations in cancer (132). A targeted sequencing of epigenetic regulators revealed that the highest frequency of epigenetic mutations occurred in high-grade glioma, ALL and medulloblastoma (133).

Recently, recurrent mutations occurring in histone genes, named “oncohistones”, have been observed in CCs. For example, oncohistones mutations are found in over 60% of all pediatric high grade glioma patients. These mutations result in global aberrant histone and altered DNA methylation pattern. These findings imply that defects of the chromatin architecture underlie pediatric glioblastoma pathogenesis (134,135). Similarly to DNA methylation changes, patterns and specific oncohistone mutations can be used for the diagnosis and classification of CNS and solid tumors (128). DNMT3a mutations are also frequent in AML patients, which results in aberrant DNA methylation patterns (136), as well as mutations in genes encoding HDAC in CL (125). Epigenetic alterations are potentially reversible and are therefore more prone to corrective therapy. Recently, the first therapies targeting epigenetic alterations have entered clinical trials. For example, HDAC inhibitors are being used in clinical trials for CL patients (128).

Underlying biological mechanisms by which early-life factors contribute to CC are poorly understood. Hence, gene–environment interaction studies are becoming increasingly central in epidemiological research proposing to investigate associations between risk factors and disease. By assessing the exposure or lifestyle factor measurements, (epi)genomics can provide a molecular history of past and current exposure events, which can potentially be used as a prospective molecular assessment of archived biospecimen that are collected in retrospective study designs (137). The application of Mendelian randomization (MR) (138) using genetic proxies has helped strengthen causal inferences in observational studies. Integrating epigenomics, genomics and exposure timing in epidemiological research would be key to understanding causal factors driving CC, with important implications in biomarker-based diagnosis, targeted therapy and prevention.

D. Relevant publications

The following book chapter represents a review work describing early-life environmental exposures including birthweight and their impact on later life health outcomes with focus on cancer in both child and adulthood. We also discuss the epigenetic mechanisms which may contribute to the association of each exposure to cancer incidence. Further research is needed to determine the extent to which epigenetic modifications mediate these links between early-life factors and tumor development.

1) Sexton-Oates A, **Novoloaca A**, Ghantous A[#], Herceg Z[#]. Cancer. *Environmental Epigenetics in Toxicology and Public Health*. Submitted.
[#]Equal contribution.

E. Methodological implications

Methylation data like many other omics can be affected by systematic variation due to technical processing of the biospecimens or the time difference in processing the samples. Throughout the thesis work, we have invested in progressively optimizing existing bioinformatic and biostatistic pipelines for methylation data. Besides investigating DNA methylation data, we also studied other types of omics on the same samples. Individual analysis of one omic allows to identify specific signals for each dataset, however integrative analysis of several omics allows the identification of a shared pathway and this is particularly important in the case of a complex disease such as cancer. Due to the recent advances in high-throughput technologies, it is critical to continuously monitor and benchmark the different statistical methods in light of the rapid evolvement of the omics field. For these reasons, the second section of this manuscript will describe three major methodology investigations performed as part of the PhD thesis.

Environmental Epigenetics in Toxicology and Public Health
Chapter 7. Cancer

Alexandra Sexton-Oates, Alexei Novoloaca, Akram Ghantous[#], and Zdenko Herceg*[,#]
Epigenetics group, International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, Lyon, France

*Correspondence: Z. Herceg. Epigenetics Group, International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, 69372 Lyon Cedex 08, France (Tel: +33 4 72 73 83 98; Fax: +33 4 72 73 83 22) E-mail: hercegz@iarc.fr
[#] Shared senior authorship

Introduction

Early-life environmental exposures, that is factors we are exposed to *in utero* and during childhood, are known to be determinants of later life health. Such exposures can be tangible such as pharmaceuticals, chemicals and metals, or infectious agents, but can also include socio-economic and psychological factors, as well as proxy exposures such as birth weight which, whilst not a direct exposure, are the manifestation of a number of known and potentially unknown environmental exposures. Later life health impacts of early-life exposures include increased risk of metabolic and cardiovascular disease, cancer, respiratory conditions and mental health disorders. In this chapter we will specifically focus on cancer.

The mechanisms by which early-life environmental exposures contribute to cancer development are relatively unstudied, but it is hypothesised that epigenetic mechanisms may be involved (Figure 1). Epigenetic mechanisms play an important role in cancer development and progression [1]. DNA methylation in particular has been extensively studied, and has been shown to be altered in almost all human cancers. In contrast to the healthy cell, tumour cells tend to display a pattern of genome-wide hypomethylation and promoter-specific hypermethylation [2]. Genome-wide hypomethylation is linked with chromosomal instability, translocations and loss of imprinting [3, 4]; and via these mechanisms, genome-wide hypomethylation has been shown both *in vitro* and *in vivo* to contribute to tumourigenesis [5-8]. Promoter-specific DNA methylation is associated with alterations in gene expression, including silencing of tumour suppressor genes and activation of oncogenes [3]. Other epigenetic mechanisms known to operate in carcinogenesis are changes to histone modifications, thereby, altering chromatin state and the regulation of gene transcription, as well as microRNA expression [1]. The precise timing of epigenetic alterations in cancer can be difficult to establish, whether they are a consequence of cancer or precede tumour development; however, it is thought to be a combination of both and may be cancer-specific [3].

Epigenetic mechanisms are also integral to embryonic development. DNA methylation profile, the most well studied epigenetic mechanism in human development, is highly regulated in the prenatal period, whereby the genome undergoes widespread demethylation shortly after fertilisation and re-methylation following implantation of the blastocyst [9]. Through a number of large epigenome-wide association studies, it is now appreciated that the prenatal environment is able to modify DNA methylation profile as measured at birth. Maternal smoking, under-nutrition, over-nutrition and even levels of individual nutritional compounds such as folate and riboflavin during pregnancy have all been shown to influence DNA methylation profile at birth [10].

The importance of epigenetic mechanisms in both cancer and prenatal development, as well as the evidence of maternal factors during pregnancy influencing epigenetic status in early infancy, suggest that epigenetics may be an important mediator of early-life exposures on later cancer development [11]. In this chapter, we provide an overview of early-life factors, grouped into themes of

pharmacological, radiation, lifestyle and personal behaviour, infectious agents, and proxy exposures, with known ties to cancer development in both child and adulthood. We focus on particularly well-characterised examples of such exposures, and discuss the epigenetic mechanisms which may contribute to the association of each exposure to cancer incidence. A summary of available evidence is shown in Table 1.

Part 1: Pharmacological Exposures

Pharmacological agents, such as hormonal contraceptives and hormone replacements, chemotherapy and immunosuppressive medicines produce a wide-array of side effects, including increasing risk of developing cancer. Here we detail the links between early-life exposure to selected hormone-based pharmaceuticals and cancer risk.

Early-life exposure to several synthetic hormone-containing medications has been shown to increase the risk of developing cancer in childhood and later in life. A well-characterised example of such an association is that of maternal use during pregnancy of Diethylstillbestrol and vaginal cancer in offspring. Evidence of a link between exposure to Diethystillbestrol (DES) prenatally and an increased risk of cancer in offspring began to emerge in the 1970s, with the publication of a small retrospective case-control study of young women with adenocarcinoma of the vagina [12]. DES, prescribed to pregnant women to prevent stillbirth and miscarriage over several decades, has subsequently been associated with an increased risk of clear cell adenocarcinoma of the vagina and cervix, as well as breast cancer, in girls and women exposed *in utero* [13-17]. There may also be an association with increased risk of testicular cancer in prenatally exposed men [18].

DES, a synthetic oestrogen, was used extensively in medical practice as an oestrogen replacement therapy and to prevent pregnancy complications. However, as an oestrogen receptor agonist it is also able to induce changes in developing reproductive tissues. Experimental models in mice have demonstrated that, when female mice are exposed to DES during gestation they develop hyperplasia, malformations, and benign and malignant neoplasms of the reproductive tract and ovaries [19-21].

The mechanism by which DES induces carcinogenesis is not fully understood; however, it is hypothesised that epigenetic mechanisms may play an important role [22]. As a synthetic oestrogen, DES has been shown to alter gene expression and DNA methylation patterns of genes normally under the control of oestrogen. For example, *HOXA10*, an important regulator of reproductive organ development exhibits altered DNA methylation and gene expression profiles in the uterine tissue of mice exposed prenatally [23]. *Lactoferrin*, also expressed in reproductive tissues and a regulator of cell growth and differentiation under the control of oestrogen, displays altered DNA methylation in the uterine tissue of mice exposed prenatally [24]. Interestingly, both genes have been shown to be altered in mice by *in utero* DES exposure only, and not as a consequence of exposure in adulthood [23, 24].

Other commonly prescribed pharmaceuticals containing female sex hormones which children can be exposed to in early life, or around conception, are hormonal contraceptives and fertility medications. Hormonal contraceptives, such as the pill and mini-pill, contain oestrogen and/or progesterone and are known to increase the risk of breast cancer in adult women [25]. There is now evidence to suggest that the use of hormonal contraceptives may increase the risk of cancer in subsequent offspring. A Danish population-based cohort study recently showed that children whose mothers took any form of hormonal contraceptive in the three months prior to pregnancy, or during pregnancy, are at increased risk of non-lymphoid leukaemia compared to those whose mother's never took hormonal

contraceptives [26]. The study estimated one leukaemia case would occur for every 50,000 children exposed, which corresponds to 25 additional cases during the nine-year study [26].

Hormone-containing fertility medications such as gonadotropins and progesterone are used in conjunction with assisted reproductive technology (ART) procedures [27]. While studies have shown no increased risk of cancer in children conceived by assisted reproductive technologies (ART) [27, 28], use of specific hormones during pregnancy have been associated with cancer in offspring. Progesterone is prescribed to prepare the embryo for implantation as part of ART treatment [29], and during pregnancy to reduce the risk of pre-term birth [30]. Children exposed *in utero* to progesterone-containing medications have been shown to be at increased risk of developing acute lymphocytic leukaemia and sympathetic nervous system tumours [27, 31].

Due to the link between hormonal contraceptive use and adult-onset breast cancer, the role of progesterone in tumourigenesis has primarily been investigated in adult women, and interestingly has different effects in different tissues. In mammary tissues, progesterone works together with oestrogen to induce stem cell expansion, whilst in ovarian and endometrial tissues it opposes the action of oestrogen to induce cell hypoplasia [32]. As the associations between *in utero* exposure to progesterone and later-onset cancer are relatively recent, the mechanism by which this may occur is currently unstudied. However, in the adult exposure setting it has been shown that progesterone activity in breast cells is able to regulate particular micro RNAs that promote growth of stem cell-like populations and therefore contribute to pre-malignant transformation of mammary cells [32, 33].

Part 2: Ionizing Radiation

Ionizing radiation, that which can cause atoms to become ionized, has long been recognised as a carcinogen. Exposure to ionizing radiation may come from various sources such as diagnostic imaging, radiation treatment for cancer, cosmic radiation, and nuclear power generation. The effects of ionizing radiation have primarily been studied on highly exposed populations, such as nuclear power plant workers, and miners, that is generally adult men. The consequences of radiation exposure on more varied populations was first able to be investigated after the atomic bombings in Japan [34]. Studies of cancer incidence in survivors of the bombing of Hiroshima and Nagasaki have shown an increased risk of developing solid cancers during child and adulthood, specifically for those exposed during early childhood (up to six years), but not during the prenatal period [35]. Interestingly, the risk of death from solid cancer in survivors increased with decreasing age at time of exposure, highlighting the importance of early-life exposures to lifetime cancer risks [36].

More commonly, early-life exposure to ionizing radiation comes from its use as a diagnostic and therapeutic tool in medicine. An association between prenatal exposure to radiation through diagnostic x-ray imaging and an increased risk of childhood cancer was first identified in 1956 in a case-control study of 1,500 children who had passed away from leukaemia in the United Kingdom [37]. Since then, there have been many further studies investigating the association (reviewed in detail in [38]) which have in contrast identified very little evidence for an association between prenatal x-ray exposure and childhood cancer, specifically leukaemia, non-Hodgkin lymphoma, and central nervous system tumours [38]. Reasons for this potential difference in findings between the 1950s and now include differences in radiation dose over time, as well as differences in study methodology. Nevertheless, the imaging community remains cautious with the use of diagnostic imaging of pregnant women and where possible other imaging techniques such as ultrasound may be used [39].

An association between childhood x-ray exposure and development of leukaemia in childhood has been demonstrated [40-44], whilst the risk of other common childhood cancers such as CNS tumours and non-Hodgkin lymphoma have not been shown to be correlated with childhood x-ray history [38]. More recently, attention has turned to the possible deleterious effects of CT scanning children, given

its increasing use worldwide [45]. It has been demonstrated in several large studies including a population-based study of 10 million children and adolescents, that the risk of many cancers including leukaemia and brain cancer is significantly increased after exposure to CT scanning in childhood, in a dose-dependent manner [46, 47]. The largest risk ratios were found for CT scans of the chest, abdomen and pelvis, with abdomen and pelvis CT scans resulting in the highest risk for leukaemia, and CT scans of the head associated with the greatest risk for CNS tumours [46].

The dose of ionizing radiation delivered during diagnostic x-ray and CT scan procedures, measured in grays (Gy) or milligrays (mGy), is rarely above 100mGy [45], whilst the cumulative dose administered during radiotherapy treatment of childhood cancers may be in excess of 50Gy. It is therefore not surprising that children are at greater risk of cancer following radiation therapy than following diagnostic imaging procedures. Children with cancer who are treated with radiotherapy have higher incidence of secondary cancers during childhood and young adulthood than both the general population, and those children treated with chemotherapy only [48-50]. Radiotherapy for childhood cancer is a risk factor for development of secondary CNS, breast, thyroid, bone and soft tissue cancers, and this risk is typically dose-dependent [51]. The association between breast cancer and radiotherapy treatment for childhood Hodgkin's lymphoma is particularly important, as the lifetime risk of developing breast cancer is greater in these patients than in women carrying the well-known a *BRCA1* mutation [52].

Ionizing radiation can promote cancer development through the creation of double-stranded DNA breaks, leading to chromosomal abnormalities and genetic mutations [34]. These can lead directly to the activation of oncogenes or the silencing of tumour suppressor genes therefore providing an environment amenable to malignant cell transformation. However, recently there has been increasing investigation of non-targeted effects, that is the effect of ionizing radiation on unexposed cells or the progeny of exposed cells [34, 53]. Epigenetic mechanisms play a role in these non-targeted effects, such as radiation induced bystander effects, and radiation induced genomic instability, which can also lead to the development of cancer [54].

Radiation induced genomic instability (RIGI), produces mutations, chromosomal instability and gene amplification in the progeny of exposed cells, and it is hypothesised that these effects may be the result of inherited epigenetic signatures from the original exposed cells. Alterations in DNA methylation profile have been demonstrated in animal and human models following exposure to ionizing radiation. For example, whole body irradiation with x-rays of rats, mice and hamsters, as well as irradiation of human cell lines, has resulted in genome-wide hypomethylation, associated with genome instability, and promoter specific hypermethylation, associated with repression of tumour suppressor genes [55-57]. It is proposed that these changes in DNA methylation can manifest in daughter cells, resulting in genomic instability and potential for malignant transformation, despite never being exposed to radiation [54].

Part 3. Lifestyle and Personal Behaviour

There are a number of personal behaviour and lifestyle risks associated with poor health outcomes; these include the types of foods we eat, activity levels, and use of substances such as alcohol, tobacco, and illicit drugs. Here we focus on the effects of early-life exposure to two known carcinogenic substances used commonly world-wide, alcohol and tobacco smoking.

Alcohol

In 2018, alcohol consumption was estimated to be responsible for 3 million deaths worldwide [58]. It is a potent carcinogen, whereby there is sufficient evidence in humans to link alcohol consumption with risk of breast, colorectal, liver, oesophageal, oral and pancreatic cancers, and it is estimated to

be responsible for 11% and 5% of worldwide colorectal and breast cancer cases respectively [58]. Early-life exposure to alcohol primarily occurs *in utero*, and is associated with a variety of negative health outcomes for offspring including low birthweight, premature birth, craniofacial dysmorphia, and neurocognitive deficits [59]. In contrast, there is little evidence of an effect on childhood health from alcohol exposure through breastfeeding, likely due to the minimal concentrations of alcohol found in breast milk [60].

There is now also evidence for an association between prenatal alcohol exposure and childhood cancer incidence. A 2017 meta-analysis of 39 case-control studies by Karalexi *et al.* found a dose-dependent relationship between maternal alcohol consumption during pregnancy and risk of acute myeloid leukaemia (AML) in childhood [61]. An earlier meta-analysis in 2010 by Latino-Martel *et al.* also identified a significant association between prenatal alcohol consumption and risk of childhood AML [62]. Interestingly neither study found an association with the most common form of childhood cancer, acute lymphoblastic leukaemia [61, 62]. With regards to other common childhood cancers, there has also been no relationship found between maternal alcohol consumption and incidence of central nervous system tumours [63] or neuroblastoma [64].

Given the risks associated with alcohol consumption during adulthood and breast cancer, it is hypothesised that there may also be a link between *in utero* alcohol exposure and breast cancer. In animal studies, prenatal alcohol exposure has been shown to increase tumorigenesis in the mammary glands of rats [65, 66]; however data in humans is scarce. Further epidemiological evidence for an association between prenatal alcohol exposure and adult cancer incidence could be gathered from studies of foetal alcohol spectrum disorder (FASD). However, as FASD is a relatively recent designation, with guidelines for diagnosis in the US first published in 1996, there is currently limited data on the health outcomes in this population in adulthood [67, 68].

There are many potential mechanisms by which alcohol contributes to tumour formation, including DNA damage, oxidative stress and inflammation, and changes to oestrogen signalling [69]. The specific carcinogenic components of alcohol in humans are ethanol and its metabolite acetaldehyde, the formation of which produces reactive oxygen species contributing to inflammation whilst acetaldehyde itself can produce mutations and chromosomal abnormalities [69]. Epigenetic modifications may also play a role in the mechanism of alcohol carcinogenesis, for example through the depletion of S-adenosyl methionine (SAMe). SAMe is known as the universal methyl group donor for DNA methyltransferases, the enzymes responsible for catalysing reactions which result in the addition of methyl groups to DNA [70]. It is produced in the liver from dietary methionine in the presence of methionine adenosyltransferase (MTA). Alcohol consumption can alter methionine metabolism, as well as reduce the absorption of dietary folate, producing a decrease in the availability of SAMe and subsequently genome-wide DNA hypomethylation and genomic instability [70, 71]. Additionally, the enzyme MTA is encoded for by two genes, one of which, *MAT1A*, has been shown to be epigenetically silenced in patients with alcohol-induced liver damage, which therefore also contributes to a reduction in SAMe availability and promotes further alterations to the DNA methylation profile and malignant cell transformation [69]. Prenatal exposure to alcohol has been shown to alter DNA methylation profiles in mice and rat pups [72, 73], and in children [74, 75], which may contribute to the associations between *in utero* alcohol exposure and later life health.

Tobacco Smoking

Smoking is associated with an extensive array of negative health effects, including cancer. The number of deaths attributed to smoking worldwide is 6 million per year [76]. Tobacco smoking is known to cause leukaemia, cervical, colorectal, kidney, lung, oesophageal, oral, ovarian, and pancreatic cancers, while second hand smoke exposure is associated with lung cancer [77]. Prenatal exposure to smoking is associated with a number of adverse outcomes including miscarriage, stillbirth, preterm birth and

low birthweight [78]. *In utero* exposure to smoking also increases the risk of cardiovascular and metabolic disease such as obesity and hypertension [79].

The associations between early-life exposure and cancer in childhood and adulthood are not as well-established as that of adult smoking; nonetheless, some positive associations have been reported. Maternal smoking during pregnancy has been found to increase the risk of childhood central nervous system tumours, particularly glioma and ependymoma, as well as retinoblastoma [80-82]. Paternal smoking prior to conception has been associated with an increased risk of childhood acute myeloid and lymphoid leukaemia, whilst paternal smoking during pregnancy and early childhood is associated with increased risk of acute myeloid leukaemia [83, 84]. With regard to adult onset cancers, women who have never smoked but were exposed to environmental tobacco smoke *in utero* or during childhood (up to 18 years of age), may be at higher risk of breast cancer development than those not exposed to environmental tobacco smoke [85], and both men and women exposed to passive smoking in childhood may have an increased risk of pancreatic cancer in adulthood [86]. Importantly however, there have been many studies investigating early-life exposure to smoking which have found no associations with childhood or adult cancer, and further evidence is required to determine the true risks.

Tobacco cigarettes contain at least fifteen compounds known to be carcinogenic to humans including polycyclic aromatic hydrocarbons, *N*-nitrosamines, aromatic amines, formaldehyde, benzene and a number of metals [87]. Consequently, the mechanism by which smoking causes cancer is likely a complex interplay of a number of different processes caused by these compounds in isolation or even in combination. General mechanisms of smoking induced tumour formation include formation of DNA adducts leading to DNA damage, mutations and chromosomal abnormalities, and receptor binding of cigarette components leading to altered signalling pathways [87].

Epigenetic alterations, specifically to the DNA methylation profile, in response to smoking are well characterised. Differences in DNA methylation profile have been identified in lung tissue and peripheral blood between smokers and non-smokers, and these differences can also be seen in gene expression profiling in many of the same genes [88, 89]. Interestingly, some genes such as *AHRR*, the aryl hydrocarbon receptor repressor which plays a role in cell proliferation and apoptosis, display a pattern of DNA methylation level in blood which reflects smoking status, whereby methylation increases in a linear fashion with length of time since quitting smoking [89]. Although *AHRR* is differentially methylated and expressed in response to smoking, whether it is also involved in the pathogenesis of smoking induced cancer is not yet clear. One study which comprised a number of different case-control cohorts identified that *AHRR* methylation is associated with tobacco smoking and the specific sites found to be differentially methylated in *AHRR* are also associated with lung cancer risk, supporting the hypothesis that *AHRR* may not just be a biomarker of tobacco smoke but may also play a role in the pathogenesis of lung cancer [90].

DNA methylation profile differences have also been investigated in children who were exposed to smoking *in utero*. A large meta-analysis identified nearly 6,000 differentially methylated sites in the blood of newborns between those whose mothers smoked during pregnancy and those whose mothers did not. Of the approximately 6,000 differentially methylated sites, 4,000 of these were also found differentially methylated with the same direction of effect later in childhood [91]. Epigenetic mechanisms may also play a role in mediating the association between paternal smoking prior to conception and childhood cancer risk. Differences in the genome-wide DNA methylation patterns in sperm between men who smoke and never-smoked have been identified, suggesting a possible mechanism by which paternal smoke could influence offspring health prior to birth [92].

Part 4. Infectious Agents

The contribution of infectious agents to cancer risk is well established, with a number of viral, bacterial and parasitic agents classified as Group 1 carcinogens by the World Health Organisation [77]. It is estimated that approximately 18% of cancers world-wide can be attributed to infectious agents, the most contributing agents being *Helicobacter pylori*, human papilloma viruses and hepatitis C and B, responsible for approximately 5% of all cancers each [93]. These agents cause lymphoma, liver, stomach, and cervical cancers in adults [93]. The epidemiology of infection-associated cancer in children differs from that of adults, where Epstein-Barr virus, human immunodeficiency virus (HIV) and human herpes virus 8 are additionally important contributors [93].

Early-life exposure to infectious agents may occur prenatally with maternal viral illness, or during childhood. The association of cancer with infectious agents has primarily been shown after childhood exposure; of note, however, is that some infections may be acquired *in utero*, such as HIV. Infection-associated cancers in children tend to be concentrated in developing countries in Africa, South America and Asia, likely due to the increased rates of infectious disease in these areas [93]. Virus-associated cancers are those where the virus is detectable within the cancer [94]. Here we highlight four infectious agents commonly acquired during infancy and early-childhood that are important risk factors for infection-associated cancers, hepatitis B, Epstein-Barr, human herpes virus 8, and HIV.

Chronic infection with hepatitis B (HBV) affects 2 billion people worldwide and is a key risk factor in the development of cirrhosis and hepatocellular carcinoma (HCC) [95]. Infants are particularly at risk of becoming chronically infected, with 90% of those infected in infancy developing a chronic infection, in contrast with 25-50% of those who acquire the infection in childhood and 5% percent in adulthood [95]. Chronic HBV infection in childhood, which predominantly occurs in Africa and Asia, is associated with an increased risk of HCC both in child and adulthood, with an estimated lifetime risk of 9-24% [95]. Importantly, public health efforts are able to reduce the number of children being diagnosed with HCC by reducing the rate of HBV infection. Following the introduction of an infant HBV vaccination schedule in Taiwan, the risk of HCC in childhood decreased [96]. In Japan in 1986 a program of administering hepatitis B immune globulins to newborn infants with HBV infected mothers was implemented. In the years following the intervention there was a significant decrease in the rate of childhood HCC incidence [97].

There are several hypothesised mechanisms by which HBV is believed to cause cancer; these include integration of HBV DNA into the hepatocyte genome resulting in chromosomal alterations and genomic instability, transcriptional activation of growth pathways through interaction with the HBx protein, and as a result of chronic inflammation or cirrhosis of the liver [98]. Although cirrhosis is found in up to 70% of cases of HCC, it is not typically a feature of childhood HCC [98]. A potential epigenetic mechanism for HBV-induced carcinogenesis is through HBx protein regulation of DNA methyltransferases. Investigations have shown HBx protein to cause upregulation of DNMT1, DNMT3A and DNMT3A2 and downregulation of DNMT3B, leading to aberrant DNA methylation profile and promotion of tumorigenesis [99].

Epstein-Barr virus in childhood is a ubiquitous and typically benign infection, yet highly associated with the development of Burkitt's and Hodgkin lymphoma. Infection generally occurs in early-life and the rate of infection varies with geographical location, for instance by the age of one, 80% of children in Uganda are likely to be positive for EBV, in contrast with 45% of children in the USA [100]. Although highly associated with development of lymphoma, not all children exposed to EBV will develop malignancy; therefore other factors play a role in EBV-associated carcinogenesis including the acquisition of oncogenic driver events [101]. EBV-positive lymphomas in childhood are most commonly found in Africa. Studies in African children have shown that approximately 90% of Burkitt's lymphoma cases, which alone make up 75% of all childhood cancers in Africa, are EBV-positive, and

over 50% of Hodgkin's lymphomas are positive [94]. This is in contrast with childhood lymphoma in the US and Europe where 30% of Burkitt's lymphoma and Hodgkin's lymphoma are EBV positive [102, 103]. The high prevalence of Burkitt's lymphoma in Africa, particularly central and southern Africa, is thought to be associated with co-infection with malaria, present at holoendemic levels in these regions [94], as well as with HIV [104]. It is hypothesised that malaria increases EBV levels whilst also supressing the immune system, allowing the proliferation of Burkitt's lymphoma precursor cells and therefore increasing the risk of malignancy [94].

EBV preferentially infects B cells, an important component of the adaptive immune system, and results in the creation of immortalised lymphoblastoid cell lines containing a latent viral genome [105]. Sequencing analyses of EBV-associated lymphomas have shown a number of genetic variations thought to contribute to cancer development after EBV infection including c-MYC translocation and *TP53* mutations; however, it has also been shown that EBV-positive tumours have greater numbers of genetic alterations than EBV-negative sporadic lymphomas [101]. This suggests additional mechanisms may be required for tumour development after EBV infection such as epigenetic changes. EBV-associated cancers have disruption of epigenetic profile, and interestingly there are distinct differences in DNA methylation profile between Burkitt's lymphoma cells carrying the EBV genome and those which do not [106]. EBV-infected B-cell lines have been shown to display decreased genome-wide DNA methylation level and a decrease in heterochromatic histone marks, which can prevent activation of tumour suppressor genes and predispose the cells to malignant transformation [106].

Human herpes virus 8 (HHV) also known as Kaposi's sarcoma-associated herpes virus (KSHV) is recognised as the cause for virtually all cases of Kaposi's sarcoma; however, as with EBV, not all those infected with KSHV will develop Kaposi's sarcoma [107]. As with EBV-associated cancers, co-infection with HIV is also an important factor in development of malignancy, the incidence of KS in the general population is 1 in 100,000, but 1 in 20 for HIV infected individuals [108]. Childhood incidence of Kaposi's sarcoma is highest in Africa where it comprises up to 50% of childhood cancer incidence whilst it is rarely seen in the US and Europe [109, 110].

The KSHV genome encodes a number of proteins with the ability to prime cells for malignant transformation. These proteins are able to inactivate tumour suppressor genes, induce angiogenesis and survival signalling, and block apoptotic signalling in infected host B and endothelial cells [111, 112]. Epigenetic mechanisms may also play a role in KSHV-mediated tumourigenesis, as KSHV is able to modify the DNA methylation profile of host genomes, promoting DNA hypermethylation by interacting with DNMTs that has been shown in *in vitro* models to increase proliferation and migration [113].

Worldwide, 37.8 million people live with human immunodeficiency virus (HIV), and 1.7 million of these are children under the age of 15 [114]. HIV is a lentivirus which damages the immune system primarily through reducing the T-cell population [115]. HIV infected individuals, both children and adults, have a higher risk than the general population of developing a number of specific cancers including Kaposi's sarcoma, non-Hodgkin's lymphoma (such as Burkitt's lymphoma) and cervical cancer. These have become known as 'AIDS-defining cancers' (ADCs) as they signify the transition from HIV to AIDS [116]. ADCs are primarily associated with co-infection of oncogenic viruses such as human herpes virus 8, EBV and HPV, and comprise up to 40% of cancer incidence in people with HIV/AIDS in contrast to approximately 5% of the general population [116, 117]. People with HIV also have higher rates of non-ADC cancers than the general population. For example, in the USA in 2010 the most common cancers diagnosed in people with HIV were non-Hodgkin's lymphoma (21% of all diagnoses) and Kaposi's sarcoma (12%), then lung (11%), anal (10%) and prostate (7%) cancer [116]. Since the introduction of highly active antiretroviral therapy (HAART), the risk of ADCs in both child and adulthood has

decreased. Importantly however, this data is primarily obtained from developed countries [116, 118-120], while the rate of non-AIDS defining malignancies in both children and adults with HIV is increasing [104, 116].

The mechanism by which HIV infection promotes tumorigenesis is not well understood; however, there are several plausible hypotheses including chronic immunosuppression allowing co-infection with oncogenic viruses (though many oncogenic viruses are ubiquitous and do not cause cancer in the majority of people infected), decreased immunosurveillance of tumour formation, and higher levels of inflammation [117]. Whilst epigenetic mechanisms have been shown to be important for integration of the virus into the host genome, little is known about the role of epigenetics in HIV-associated cancer development [121].

Part 5. Proxies for Early-Life Environmental Exposures: Birthweight

Birthweight is a marker of prenatal growth and can be used as a proxy of cumulative effects of in utero exposures on later in life health outcomes, including cancer. The potential links between birth characteristics, specifically birthweight, and childhood cancer were first suggested in the early 60s by MacMahon and Newill [122]. Subsequently, this putative association has been studied using retrospective case-control studies.

Large meta-analyses of such retrospective case-control studies have shown an association between high birthweight and an increased risk of childhood acute lymphoblastic leukaemia, neuroblastoma and CNS tumours (specifically astrocytoma) [123-126]. Low or very low birthweight has also been associated with an increased risk of childhood hepatoblastoma [127]. The majority of evidence presented here so far is based on retrospective studies, with likelihood of recall and control selection bias. As childhood cancer is rare, well-powered prospective studies have been limited. More recently, the largest birth cohort study to date encompassing 112,781 live births (including 377 cancer cases) from six geographically diverse cohorts by the International Childhood Cancer Cohort Consortium (I4C) showed that high birthweight increases the risk of childhood leukaemias and childhood cancer overall; high birthweight was also positively associated with non-leukaemia cancer among children diagnosed at age ≥ 3 years but not at younger age [128]. Overall, based on both retrospective and prospective evidence, birthweight represents one of the earliest predisposing phenotypes of childhood cancer.

Additionally, birthweight has been shown to be associated with outcomes occurring much later in life, including adult cancers (Figure 1). Several studies including large meta-analyses have identified an association between high birthweight and both pre- and postmenopausal breast cancer [129-131]; however, some have reported no significant associations [132, 133]. Some investigations have also incorporated intermediate-age growth measures in their analysis given the long latency period between early-life exposure and adult cancer development. For instance, birth weight and risk of postmenopausal breast cancer is mediated to some extent by childhood or adolescent growth, especially by adult height [131]. In premenopausal breast cancer, however, the effect of birthweight on cancer risk has been shown to be only slightly reduced after simultaneous adjustment for height and body mass index (BMI) at age 2 years and height and BMI velocities throughout childhood and adolescence [134].

Based on these studies, it is not clear accordingly whether postnatal growth underlies the pathways linking birthweight to breast cancer risk, and this could be further complicated through effect modification by menopausal status. The mechanism proposed by the World Cancer Research Fund (WCRF) is that of a long-term programming of hormonal systems, including putative roles of Insulin-Like Growth Factor (IGF)-1, the main mediator of growth hormone activity, and adipose-tissue derived oestrogen [135]. The action of both oestrogen and IGF-1 are thought to play a synergistic role in breast

carcinogenesis and in foetal growth and mammary gland development [135]. Another suggested mechanism is that the likelihood of breast cancer depends on the number of cells at risk, which is partially determined early in life [136]. Recently, a Mendelian Randomization study of genetic data from 122,977 breast cancer cases and 105,974 controls highlighted that genetic proxies of higher birthweight were not associated with an increased risk of adult breast cancer [137], suggesting that the association between birthweight and breast cancer risk may not be causal. Other adult cancers whose incidence has been variably linked with birthweight include melanoma [138], prostate [139-141], colorectal [142-145], testicular [146], kidney [147] and lung cancer [145, 147].

All these studies point out the importance of birth characteristics and of the foetal life, directly or indirectly, on the risk of different cancer types. To add a new dimension to the value of the current epidemiological studies, the next step and missing link is to investigate the molecular mechanisms underlying the associations between birthweight and cancer. One hypothesised mechanism by which birthweight may be associated with cancer is through DNA methylation. Good evidence for this was recently shown from the Pregnancy And Childhood Epigenetics (PACE) Consortium from almost 9000 neonates from 24 birth cohorts worldwide, that identified thousands of differentially methylated markers in association with birthweight, among which 1.3% persist in childhood and adolescence [148].

Conclusion

Exposure to a wide variety of environmental factors *in utero* and during childhood are associated with the development of many adverse health outcomes including cancer. Many of the associations presented in this chapter are relatively new discoveries and most still require further investigation to determine more concretely their carcinogenic status, as well as to understand the mechanisms behind their association with cancer development [149]. Importantly, the long latency period between early-life exposure and later onset of cancer, particularly in adults, makes the identification of associated links and mechanisms difficult. In summary, it is evident that many environmental exposures do not only have an effect around the time of exposure, but that early-life exposures, even prior to conception, have a potential effect upon cancer development throughout the life course.

Such long-lasting effects from relatively short exposure windows at very young ages may be the result of epigenetic changes at the time of exposure which persist throughout life, creating a cellular and genetic environment primed for tumour development. However further research is required to determine the extent to which epigenetic modifications mediate these links between early-life exposure and tumour development.

Disclaimer

The authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer / World Health Organization.

Table 1. Early life exposure and cancer risk later in life: evidence in humans

Early-Life Exposure	Early-Life Cancer	Adult Cancer	Potential Epigenetic Mechanism	Ref
Pharmacological Diethylstillbestrol	Vagina and cervix	Vagina, cervix, breast, and testicular	Altered DNA methylation (and expression) of reproductive regulator genes *HOXA10* and *Lactoferrin* in mice uterine tissues	12-24
Pharmacological Hormonal contraceptives	Blood (non-lymphoid)	Breast	Altered regulation of microRNAs which promote growth of stem cell-like populations in mammary cell lines	25-26, 32-33
Pharmacological Progesterone	Blood (non-lymphoid) and sympathetic nervous system		Altered regulation of microRNAs which promote growth of stem cell-like populations in mammary cell lines	27, 31-33
Ionizing Radiation CT scanning, radiotherapy	Blood, central nervous system, breast, thyroid, bone and soft tissue	Blood, central nervous system, breast, thyroid, bone and soft tissue	Genome-wide hypomethylation leading to genomic instability, and promoter-specific hypermethylation in rats, mice, hamsters and human cell lines	46-57
Lifestyle Alcohol	Blood (non-lymphoid)		Depletion of the methyl group donor, SAM and epigenetic silencing of methionine adenosyltransferase leading to genome-wide hypomethylation and genomic instability Altered DNA methylation profile in mice and rats	61-62, 69-75
Lifestyle Tobacco smoking	Central nervous system, retinoblastoma, blood	Breast and pancreas	*AHRR* methylation Genome-wide DNA methylation alterations in sperm	80-92
Infections Hepatitis B Virus	Hepatocellular carcinoma	Hepatocellular carcinoma	Upregulation of DNMT1, DNMT3A and DNMT3A2 and downregulation of DNMT3B	95-97, 99
Infections Epstein-Barr Virus	Burkitt's and Hodgkin's lymphoma	Burkitt's and Hodgkin's lymphoma	Genome-wide hypomethylation level and a decrease in heterochromatic histone marks	94, 100-101, 104, 106
Infections Human Herpes Virus 8	Kaposi's sarcoma	Kaposi's sarcoma	DNA hypermethylation by interacting with DNMTs	107-113
Infections	Kaposi's sarcoma, non-	Kaposi's sarcoma, non-		114-117

Human Immunodeficiency Virus	Hodgkin's lymphoma, and cervical	Hodgkin's lymphoma, and cervical		

* Abbreviations: SAM (S-Adenosyl Methionine); DNMT (DNA Methyltransferase

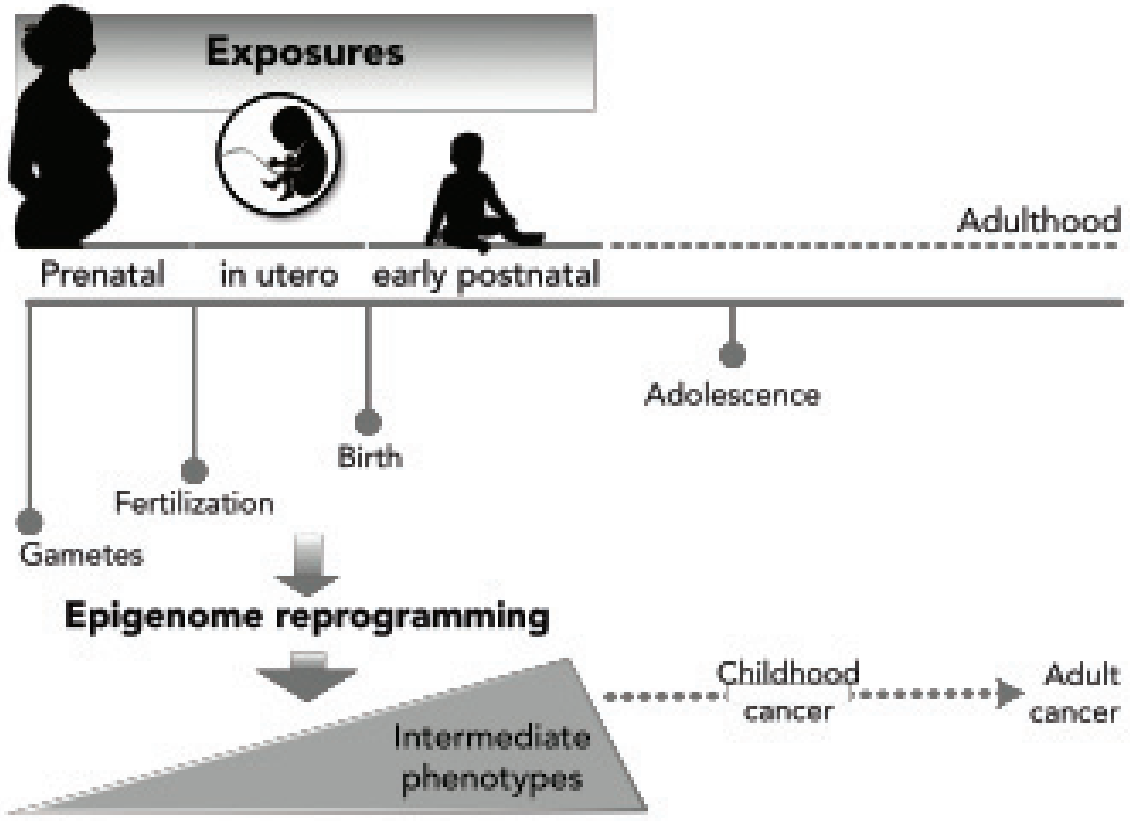


Figure 1. In utero and early life exposures and the concept of a developmental origin of cancer in childhood and later life. This concept postulates that susceptibility to childhood and adult diseases (including cancer) may be conveyed through epigenetic changes induced by exposures during in utero and early postnatal life. Different in utero/early life exposures (including pharmacological agents, lifestyle factors, chemicals/metals, infectious agents, radiation, socioeconomic status, psychological factors) may induce epigenome reprogramming that in turn may deregulate stem/progenitor cells, hormonal and growth pathways and metabolic changes that may influence intermediate phenotypes (such as birthweight). These deregulated processes may be propagated though aberrant epigenome states throughout the life-course potentially constituting the susceptibility to diseases (including cancer) in later life.

REFERENCES

1. Feinberg, A.P., *Phenotypic plasticity and the epigenetics of human disease.* Nature, 2007. 447(7143): p. 433.
2. Esteller, M., *Epigenetics in cancer.* New England Journal of Medicine, 2008. 358(11): p. 1148-1159.
3. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nature reviews genetics, 2006. 7(1): p. 21.

4. Feinberg, A.P. and B. Tycko, *The history of cancer epigenetics.* Nature Reviews Cancer, 2004. 4(2): p. 143.
5. Schmid, M., T. Haaf, and D. Grunert, *5-Azacytidine-induced undercondensations in human chromosomes.* Human genetics, 1984. 67(3): p. 257-263.
6. Eden, A., et al., *Chromosomal instability and tumors promoted by DNA hypomethylation.* Science, 2003. 300(5618): p. 455-455.
7. Gaudet, F., et al., *Induction of tumors in mice by genomic hypomethylation.* Science, 2003. 300(5618): p. 489-492.
8. Holm, T.M., et al., *Global loss of imprinting leads to widespread tumorigenesis in adult mice.* Cancer cell, 2005. 8(4): p. 275-285.
9. Smallwood, S.A. and G. Kelsey, *De novo DNA methylation: a germ cell perspective.* Trends Genet, 2012. 28(1): p. 33-42.
10. Fernandez-Twinn, D.S., et al., *Intrauterine programming of obesity and type 2 diabetes.* Diabetologia, 2019: p. 1-13.
11. Herceg, Z., et al., *Roadmap for investigating epigenome deregulation and environmental origins of cancer.* International journal of cancer, 2018. 142(5): p. 874-882.
12. Herbst, A.L., H. Ulfelder, and D.C. Poskanzer, *Adenocarcinoma of the vagina: association of maternal stilbestrol therapy with tumor appearance in young women.* New England journal of medicine, 1971. 284(16): p. 878-881.
13. Verloop, J., et al., *Cancer risk in DES daughters.* Cancer Causes & Control, 2010. 21(7): p. 999-1007.
14. Troisi, R., et al., *Cancer risk in women prenatally exposed to diethylstilbestrol.* International journal of cancer, 2007. 121(2): p. 356-360.
15. Hoover, R.N., et al., *Adverse health outcomes in women exposed in utero to diethylstilbestrol.* New England Journal of Medicine, 2011. 365(14): p. 1304-1314.
16. Strohsnitter, W.C., et al., *Cancer risk in men exposed in utero to diethylstilbestrol.* Journal of the National Cancer Institute, 2001. 93(7): p. 545-551.
17. Troisi, R., E.E. Hatch, and L. Titus, *The diethylstilbestrol legacy: a powerful case against intervention in uncomplicated pregnancy.* Pediatrics, 2016. 138(Supplement 1): p. S42-S44.
18. Palmer, J.R., et al., *Prenatal diethylstilbestrol exposure and risk of breast cancer.* Cancer Epidemiology and Prevention Biomarkers, 2006. 15(8): p. 1509-1514.
19. McLachlan, J.A., R.R. Newbold, and B.C. Bullock, *Long-term effects on the female mouse genital tract associated with prenatal exposure to diethylstilbestrol.* Cancer Research, 1980. 40(11): p. 3988-3999.
20. Newbold, R.R., B.C. Bullock, and J.A. McLachlan, *Uterine adenocarcinoma in mice following developmental treatment with estrogens: a model for hormonal carcinogenesis.* Cancer research, 1990. 50(23): p. 7677-7681.
21. Newbold, R.R., A.B. Moore, and D. Dixon, *Characterization of uterine leiomyomas in CD-1 mice following developmental exposure to diethylstilbestrol (DES).* Toxicologic pathology, 2002. 30(5): p. 611-616.
22. Li, S., et al., *Environmental exposure, DNA methylation, and gene regulation: lessons from diethylstilbesterol-induced cancers.* Annals of the New York Academy of Sciences, 2003. 983(1): p. 161-169.
23. Bromer, J.G., et al., *Hypermethylation of homeobox A10 by in utero diethylstilbestrol exposure: an epigenetic mechanism for altered developmental programming.* Endocrinology, 2009. 150(7): p. 3376-3382.
24. Li, S., et al., *Developmental exposure to diethylstilbestrol elicits demethylation of estrogen-responsive lactoferrin gene in mouse uterus.* Cancer research, 1997. 57(19): p. 4356-4359.
25. Mørch, L.S., et al., *Contemporary hormonal contraception and the risk of breast cancer.* New England Journal of Medicine, 2017. 377(23): p. 2228-2239.

26. Hargreave, M., et al., *Maternal use of hormonal contraception and risk of childhood leukaemia: a nationwide, population-based cohort study.* The Lancet Oncology, 2018. 19(10): p. 1307-1314.
27. Hargreave, M., et al., *Maternal use of fertility drugs and risk of cancer in children—A nationwide population-based cohort study in D enmark.* International Journal of Cancer, 2015. 136(8): p. 1931-1939.
28. Klip, H., et al., *Risk of cancer in the offspring of women who underwent ovarian stimulation for IVF.* Human Reproduction, 2001. 16(11): p. 2451-2458.
29. van der Linden, M., et al., *Luteal phase support for assisted reproduction cycles.* Cochrane Database of Systematic Reviews, 2015(7).
30. Daskalakis, G., et al., *Prevention of spontaneous preterm birth.* Archives of gynecology and obstetrics, 2019. 299(5): p. 1261-1273.
31. Mandel, M., et al., *Hormonal treatment in pregnancy: a possible risk factor for neuroblastoma.* Medical and pediatric oncology, 1994. 23(2): p. 133-135.
32. Diep, C.H., et al., *Progesterone action in breast, uterine, and ovarian cancers.* Journal of molecular endocrinology, 2015. 54(2): p. R31.
33. Horwitz, K.B. and C.A. Sartorius, *Progestins in hormone replacement therapies reactivate cancer stem cells in women with preexisting breast cancers: a hypothesis.* The Journal of Clinical Endocrinology & Metabolism, 2008. 93(9): p. 3295-3298.
34. Little, J.B., *Radiation carcinogenesis.* Carcinogenesis, 2000. 21(3): p. 397-404.
35. Preston, D.L., et al., *Solid cancer incidence in atomic bomb survivors exposed in utero or as young children.* Journal of the National Cancer Institute, 2008. 100(6): p. 428-436.
36. Ozasa, K., et al., *Studies of the mortality of atomic bomb survivors, Report 14, 1950–2003: an overview of cancer and noncancer diseases.* Radiation research, 2011. 177(3): p. 229-243.
37. Stewart, A., et al., *MALIGNANT DISEASE IN CHILDHOOD AND DIAGNOSTIC IRRADIATION IN UTERO.* The Lancet, 1956. 268(6940): p. 447.
38. Schulze-Rath, R., G.P. Hammer, and M. Blettner, *Are pre-or postnatal diagnostic X-rays a risk factor for childhood cancer? A systematic review.* Radiation and environmental biophysics, 2008. 47(3): p. 301.
39. McCollough, C.H., et al., *Radiation exposure and pregnancy: when should we be concerned?* Radiographics, 2007. 27(4): p. 909-917.
40. Bartley, K., et al., *Diagnostic X-rays and risk of childhood leukaemia.* International journal of epidemiology, 2010. 39(6): p. 1628-1637.
41. Shu, X., et al., *Diagnostic X-ray and ultrasound exposure and risk of childhood cancer.* British journal of cancer, 1994. 70(3): p. 531.
42. Shu, X.O., et al., *Diagnostic X-rays and ultrasound exposure and risk of childhood acute lymphoblastic leukemia by immunophenotype.* Cancer Epidemiology and Prevention Biomarkers, 2002. 11(2): p. 177-185.
43. Infante-Rivard, C., *Diagnostic x rays, DNA repair genes and childhood acute lymphoblastic leukemia.* Health physics, 2003. 85(1): p. 60-64.
44. Infante-Rivard, C., G. Mathonnet, and D. Sinnett, *Risk of childhood leukemia associated with diagnostic irradiation and polymorphisms in DNA repair genes.* Environmental health perspectives, 2000. 108(6): p. 495-498.
45. Brenner, D.J. and E.J. Hall, *Computed tomography—an increasing source of radiation exposure.* New England Journal of Medicine, 2007. 357(22): p. 2277-2284.
46. Mathews, J.D., et al., *Cancer risk in 680 000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians.* Bmj, 2013. 346: p. f2360.
47. Pearce, M.S., et al., *Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study.* The Lancet, 2012. 380(9840): p. 499-505.

48. Harbron, R.W., et al., *Secondary malignant neoplasms following radiotherapy for primary cancer in children and young adults.* Pediatric hematology and oncology, 2014. 31(3): p. 259-267.
49. Armstrong, G.T., et al., *Long-term outcomes among adult survivors of childhood central nervous system malignancies in the Childhood Cancer Survivor Study.* JNCI: Journal of the National Cancer Institute, 2009. 101(13): p. 946-958.
50. Inskip, P.D., et al., *Radiation dose and breast cancer risk in the childhood cancer survivor study.* Journal of clinical oncology, 2009. 27(24): p. 3901.
51. Bhatia, S. and C. Sklar, *Second cancers in survivors of childhood cancer.* Nature Reviews Cancer, 2002. 2(2): p. 124.
52. Moskowitz, C.S., et al., *Breast cancer after chest radiation therapy for childhood cancer.* Journal of Clinical oncology, 2014. 32(21): p. 2217.
53. Morgan, W.F. and M.B. Sowa, *Non-targeted effects induced by ionizing radiation: mechanisms and potential impact on radiation induced health effects.* Cancer letters, 2015. 356(1): p. 17-21.
54. Aypar, U., W.F. Morgan, and J.E. Baulch, *Radiation-induced genomic instability: are epigenetic mechanisms the missing link?* International journal of radiation biology, 2011. 87(2): p. 179-191.
55. Pogribny, I., et al., *Fractionated low-dose radiation exposure leads to accumulation of DNA damage and profound alterations in DNA and histone methylation in the murine thymus.* Molecular cancer research, 2005. 3(10): p. 553-561.
56. Pogribny, I., et al., *Dose-dependence, sex-and tissue-specificity, and persistence of radiation-induced genomic DNA methylation changes.* Biochemical and biophysical research communications, 2004. 320(4): p. 1253-1261.
57. Kalinich, J.F., G.N. Catravas, and S.L. Snyder, *The effect of γ radiation on DNA methylation.* Radiation research, 1989. 117(2): p. 185-197.
58. *Global Status Report on Alcohol and Health, 2018*, V.P.a.D. Rekve, Editor. 2018, World Health Organisation Geneva.
59. Chiodo, L.M., et al., *Prenatal Alcohol Screening During Pregnancy by Midwives and Nurses.* Alcoholism: Clinical and Experimental Research, 2019. 43(8): p. 1747-1758.
60. Haastrup, M.B., A. Pottegard, and P. Damkier, *Alcohol and breastfeeding.* Basic Clin Pharmacol Toxicol, 2014. 114(2): p. 168-73.
61. Karalexi, M.A., et al., *Parental alcohol consumption and risk of leukemia in the offspring: a systematic review and meta-analysis.* Eur J Cancer Prev, 2017. 26(5): p. 433-441.
62. Latino-Martel, P., et al., *Maternal alcohol consumption during pregnancy and risk of childhood leukemia: systematic review and meta-analysis.* Cancer Epidemiol Biomarkers Prev, 2010. 19(5): p. 1238-60.
63. Bailey, H.D., et al., *Parental smoking, maternal alcohol, coffee and tea consumption and the risk of childhood brain tumours: the ESTELLE and ESCALE studies (SFCE, France).* Cancer Causes & Control, 2017. 28(7): p. 719-732.
64. Rios, P., et al., *Parental smoking, maternal alcohol consumption during pregnancy and the risk of neuroblastoma in children. A pooled analysis of the ESCALE and ESTELLE French studies.* International Journal of Cancer, 2019.
65. Hilakivi-Clarke, L., et al., *In utero alcohol exposure increases mammary tumorigenesis in rats.* British Journal of Cancer, 2004. 90(11): p. 2225-2231.
66. Polanco, T.A., et al., *Fetal alcohol exposure increases mammary tumor susceptibility and alters tumor phenotype in rats.* Alcoholism: Clinical and Experimental Research, 2010. 34(11): p. 1879-1887.
67. Stratton, K., C. Howe, and F.C. Battaglia, *Fetal alcohol syndrome: Diagnosis, epidemiology, prevention, and treatment*. 1996: National Academies Press.

68. Moore, E.M. and E.P. Riley, *What happens when children with fetal alcohol spectrum disorders become adults?* Current developmental disorders reports, 2015. 2(3): p. 219-227.
69. Seitz, H.K. and F. Stickel, *Molecular mechanisms of alcohol-mediated carcinogenesis.* Nature Reviews Cancer, 2007. 7(8): p. 599.
70. Varela-Rey, M., et al., *Alcohol, DNA methylation, and cancer.* Alcohol research: current reviews, 2013. 35(1): p. 25.
71. Mead, E.A. and D.K. Sarkar, *Fetal alcohol spectrum disorders and their transmission through genetic and epigenetic mechanisms.* 2014. 5.
72. Lussier, A.A., et al., *Prenatal Alcohol Exposure: Profiling Developmental DNA Methylation Patterns in Central and Peripheral Tissues.* Frontiers in Genetics, 2018. 9.
73. Liu, Y., et al., *Alcohol exposure alters DNA methylation profiles in mouse embryos at early neurulation.* Epigenetics, 2009. 4(7): p. 500-511.
74. Laufer, B.I., et al., *Associative DNA methylation changes in children with prenatal alcohol exposure.* Epigenomics, 2015. 7(8): p. 1259-1274.
75. Portales-Casamar, E., et al., *DNA methylation signature of human fetal alcohol spectrum disorder.* Epigenetics & chromatin, 2016. 9(1): p. 25.
76. Collaborators, G.B.D.T., *Smoking prevalence and attributable disease burden in 195 countries and territories, 1990-2015: a systematic analysis from the Global Burden of Disease Study 2015.* Lancet (London, England), 2017. 389(10082): p. 1885-1906.
77. Cogliano, V.J., et al., *Preventable exposures associated with human cancers.* Journal of the National Cancer Institute, 2011. 103(24): p. 1827-1839.
78. Cnattingius, S., *The epidemiology of smoking during pregnancy: Smoking prevalence, maternal characteristics, and pregnancy outcomes.* Nicotine & Tobacco Research, 2004. 6: p. 125-140.
79. Cupul-Uicab, L.A., et al., *In utero exposure to maternal tobacco smoke and subsequent obesity, hypertension, and gestational diabetes among women in the MoBa cohort.* Environ Health Perspect, 2012. 120(3): p. 355-60.
80. Heck, J.E., et al., *Smoking in pregnancy and risk of cancer among young children: A population-based study.* Int J Cancer, 2016. 139(3): p. 613-6.
81. Schüz, J., et al., *Risk factors for pediatric tumors of the central nervous system: Results from a German population-based case-control study.* Medical and Pediatric Oncology: The Official Journal of SIOP—International Society of Pediatric Oncology (Societé Internationale d'Oncologie Pédiatrique, 2001. 36(2): p. 274-282.
82. Brooks, D.R., et al., *Maternal smoking during pregnancy and risk of brain tumors in the offspring. A prospective study of 1.4 million Swedish births.* Cancer Causes & Control, 2004. 15(10): p. 997-1005.
83. Metayer, C., et al., *Parental Tobacco Smoking and Acute Myeloid Leukemia.* American Journal of Epidemiology, 2016. 184(4): p. 261-273.
84. Shu, X.-O., et al., *Parental alcohol consumption, cigarette smoking, and risk of infant leukemia: a Childrens Cancer Group study.* JNCI: Journal of the National Cancer Institute, 1996. 88(1): p. 24-31.
85. White, A.J., et al., *Breast cancer and exposure to tobacco smoke during potential windows of susceptibility.* Cancer Causes Control, 2017. 28(7): p. 667-675.
86. Chuang, S.-C., et al., *Exposure to environmental tobacco smoke in childhood and incidence of cancer in adulthood in never smokers in the European prospective investigation into cancer and nutrition.* Cancer Causes & Control, 2011. 22(3): p. 487-494.
87. Hecht, S.S., *Cigarette smoking: cancer risks, carcinogens, and mechanisms.* Langenbecks Arch Surg, 2006. 391(6): p. 603-13.
88. Joehanes, R., et al., *Epigenetic Signatures of Cigarette Smoking.* Circ Cardiovasc Genet, 2016. 9(5): p. 436-447.

89. Shenker, N.S., et al., *Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking.* Human molecular genetics, 2012. 22(5): p. 843-851.
90. Fasanelli, F., et al., *Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts.* Nature communications, 2015. 6: p. 10192-10192.
91. Bonnie, et al., *DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis.* The American Journal of Human Genetics, 2016. 98(4): p. 680-696.
92. Jenkins, T.G., et al., *Cigarette smoking significantly alters sperm DNA methylation patterns.* Andrology, 2017. 5(6): p. 1089-1099.
93. Parkin, D.M., *The global health burden of infection-associated cancers in the year 2002.* International journal of cancer, 2006. 118(12): p. 3030-3044.
94. Cader, F.Z., et al., *The contribution of the Epstein-Barr virus to the pathogenesis of childhood lymphomas.* Cancer Treat Rev, 2010. 36(4): p. 348-53.
95. Della Corte, C., et al., *Management of chronic hepatitis B in children: An unresolved issue.* Journal of Gastroenterology and Hepatology, 2014. 29(5): p. 912-919.
96. Chang, M.-H., et al., *Decreased incidence of hepatocellular carcinoma in hepatitis B vaccinees: a 20-year follow-up study.* Journal of the National Cancer Institute, 2009. 101(19): p. 1348-1355.
97. Tajiri, H., et al., *Reduction of hepatocellular carcinoma in childhood after introduction of selective vaccination against hepatitis B virus for infants born to HBV carrier mothers.* 2011. 22(3): p. 523-527.
98. Tarocchi, M., et al., *Molecular mechanism of hepatitis B virus-induced hepatocarcinogenesis.* World journal of gastroenterology, 2014. 20(33): p. 11630-11640.
99. Park, I.Y., et al., *Aberrant epigenetic modifications in hepatocarcinogenesis induced by hepatitis B virus X protein.* Gastroenterology, 2007. 132(4): p. 1476-94.
100. Hsu, J.L. and S.L. Glaser, *Epstein–Barr virus-associated malignancies: epidemiologic patterns and etiologic implications.* Critical reviews in oncology/hematology, 2000. 34(1): p. 27-53.
101. Hernandez-Vargas, H., et al., *Viral driven epigenetic events alter the expression of cancer-related genes in Epstein-Barr-virus naturally infected Burkitt lymphoma cell lines.* Scientific reports, 2017. 7(1): p. 5852-5852.
102. Orem, J., et al., *Burkitt's lymphoma in Africa, a review of the epidemiology and etiology.* Afr Health Sci, 2007. 7(3): p. 166-75.
103. Geng, L. and X. Wang, *Epstein-Barr Virus-associated lymphoproliferative disorders: experimental and clinical developments.* Int J Clin Exp Med, 2015. 8(9): p. 14656-71.
104. Chiappini, E., et al., *Pediatric Human Immunodeficiency Virus infection and cancer in the Highly Active Antiretroviral Treatment (HAART) era.* Cancer Letters, 2014. 347(1): p. 38-45.
105. Young, L.S. and A.B. Rickinson, *Epstein-Barr virus: 40 years on.* Nat Rev Cancer, 2004. 4(10): p. 757-68.
106. Niller, H.H., et al., *Role of epigenetics in EBV regulation and pathogenesis.* Future Microbiol, 2014. 9(6): p. 747-56.
107. Jackson, C.C., et al., *Kaposi Sarcoma of Childhood: Inborn or Acquired Immunodeficiency to Oncogenic HHV-8.* Pediatric Blood & Cancer, 2016. 63(3): p. 392-397.
108. Mesri, E.A., E. Cesarman, and C. Boshoff, *Kaposi's sarcoma and its associated herpesvirus.* Nature Reviews Cancer, 2010. 10(10): p. 707-719.
109. Magrath, I., et al., *Paediatric cancer in low-income and middle-income countries.* Lancet Oncol, 2013. 14(3): p. e104-16.
110. Mutyaba, I., et al., *Presentation and Outcomes of Childhood Cancer Patients at Uganda Cancer Institute.* Glob Pediatr Health, 2019. 6: p. 2333794x19849749.
111. Wen, K.W. and B. Damania, *Kaposi sarcoma-associated herpesvirus (KSHV): molecular biology and oncogenesis.* Cancer letters, 2010. 289(2): p. 140-150.

112. Kuss-Duerkop, S., J. Westrich, and D. Pyeon, *DNA Tumor Virus Regulation of Host DNA Methylation and Its Implications for Immune Evasion and Oncogenesis.* Viruses, 2018. 10(2): p. 82.
113. Wu, J., et al., *Kaposi's sarcoma-associated herpesvirus (KSHV) vIL-6 promotes cell proliferation and migration by upregulating DNMT1 via STAT3 activation.* PloS one, 2014. 9(3): p. e93478.
114. HIV/AIDS, J.U.N.P.o., *UNAIDS Data 2019*. 2019: Geneva.
115. McCune, J.M., *The dynamics of CD4+ T-cell depletion in HIV disease.* Nature, 2001. 410(6831): p. 974-979.
116. Shiels, M.S. and E.A. Engels, *Evolving epidemiology of HIV-associated malignancies.* Curr Opin HIV AIDS, 2017. 12(1): p. 6-11.
117. Ji, Y. and H. Lu, *Malignancies in HIV-Infected and AIDS Patients*, in *Infectious Agents Associated Cancers: Epidemiology and Molecular Biology*, Q. Cai, Z. Yuan, and K. Lan, Editors. 2017, Springer Singapore: Singapore. p. 167-179.
118. Kest, H., et al., *Malignancy in perinatally human immunodeficiency virus-infected children in the United States.* Pediatr Infect Dis J, 2005. 24(3): p. 237-42.
119. Simard, E.P., et al., *Long-term cancer risk among people diagnosed with AIDS during childhood.* Cancer Epidemiol Biomarkers Prev, 2012. 21(1): p. 148-54.
120. Bohlius, J., et al., *Incidence of AIDS-defining and Other Cancers in HIV-positive Children in South Africa: Record Linkage Study.* Pediatr Infect Dis J, 2016. 35(6): p. e164-70.
121. Ay, E., et al., *Epigenetics of HIV infection: promising research areas and implications for therapy.* AIDS Rev, 2013. 15(3): p. 181-8.
122. MacMahon, B. and V.A. Newill, *Birth characteristics of children dying of malignant neoplasms.* Journal of the National Cancer Institute, 1962. 28(1): p. 231-244.
123. Milne, E., et al., *Fetal growth and childhood acute lymphoblastic leukemia: findings from the childhood leukemia international consortium.* International journal of cancer, 2013. 133(12): p. 2968-2979.
124. Harder, T., A. Plagemann, and A. Harder, *Birth weight and risk of neuroblastoma: a meta-analysis.* International journal of epidemiology, 2010. 39(3): p. 746-756.
125. Harder, T., A. Plagemann, and A. Harder, *Birth weight and subsequent risk of childhood primary brain tumors: a meta-analysis.* American journal of epidemiology, 2008. 168(4): p. 366-373.
126. Dahlhaus, A., et al., *Birth weight and subsequent risk of childhood primary brain tumors: An updated meta-analysis.* Pediatric blood & cancer, 2017. 64(5): p. e26299.
127. Heck, J.E., et al., *Case-control study of birth characteristics and the risk of hepatoblastoma.* Cancer epidemiology, 2013. 37(4): p. 390-395.
128. Paltiel, O., et al., *Birthweight and Childhood Cancer: Preliminary Findings from the I nternational C hildhood C ancer C ohort C onsortium (I4C).* Paediatric and perinatal epidemiology, 2015. 29(4): p. 335-345.
129. Park, S.K., et al., *Intrauterine environments and breast cancer risk: meta-analysis and systematic review.* Breast Cancer Research, 2008. 10(1): p. R8.
130. Xue, F. and K.B. Michels, *Intrauterine factors and risk of breast cancer: a systematic review and meta-analysis of current evidence.* The lancet oncology, 2007. 8(12): p. 1088-1100.
131. Luo, J., et al., *Birth weight, weight over the adult life course, and risk of breast cancer.* International journal of cancer, 2019.
132. Andersen, Z.J., et al., *Birth weight, childhood body mass index, and height in relation to mammographic density and breast cancer: a register-based cohort study.* Breast Cancer Research, 2014. 16(1): p. R4.
133. Goldberg, M., et al., *Early-life Growth and Benign Breast Disease.* American journal of epidemiology, 2019.

134. dos Santos Silva, I., et al., *Is the association of birth weight with premenopausal breast cancer risk mediated through childhood growth?* British journal of cancer, 2004. 91(3): p. 519.
135. Research, W.C.R.F.W.A.I.f.C., *Diet, Nutrition, Physical Activity and Cancer: a Global Perspective. Continuous Update Project Expert Third Report* 2018.
136. Adami, H.-O., L.B. Signorello, and D. Trichopoulos. *Towards an understanding of breast cancer etiology*. in *Seminars in cancer biology*. 1998. Elsevier.
137. Kar, S.P., et al., *The association between weight at birth and breast cancer risk revisited using Mendelian randomisation.* European journal of epidemiology, 2019. 34(6): p. 591-600.
138. O'Rorke, M., et al., *Do perinatal and early life exposures influence the risk of malignant melanoma? A Northern Ireland birth cohort analysis.* European Journal of Cancer, 2013. 49(5): p. 1109-1116.
139. Zhou, C.K., et al., *Is birthweight associated with total and aggressive/lethal prostate cancer risks? A systematic review and meta-analysis.* British journal of cancer, 2016. 114(7): p. 839.
140. Eriksson, M., et al., *The impact of birth weight on prostate cancer incidence and mortality in a population-based study of men born in 1913 and followed up from 50 to 85 years of age.* The Prostate, 2007. 67(11): p. 1247-1254.
141. Lahmann, P., et al., *Measures of birth size in relation to risk of prostate cancer: the Malmö Diet and Cancer Study, Sweden.* Journal of developmental origins of health and disease, 2012. 3(6): p. 442-449.
142. Wang, P., et al., *Birth weight and risk of colorectal cancer: a meta-analysis.* International journal of colorectal disease, 2014. 29(8): p. 1017-1018.
143. Sandhu, M.S., et al., *Self-reported birth weight and subsequent risk of colorectal cancer.* Cancer Epidemiology and Prevention Biomarkers, 2002. 11(9): p. 935-938.
144. Nilsen, T.I., et al., *Birth size and colorectal cancer risk: a prospective population based study.* Gut, 2005. 54(12): p. 1728-1732.
145. Spracklen, C.N., et al., *Birth weight and subsequent risk of cancer.* Cancer epidemiology, 2014. 38(5): p. 538-543.
146. Michos, A., F. Xue, and K.B. Michels, *Birth weight and the risk of testicular cancer: a meta-analysis.* International journal of cancer, 2007. 121(5): p. 1123-1131.
147. Ahlgren, M., et al., *Birth weight and risk of cancer.* Cancer, 2007. 110(2): p. 412-419.
148. Küpers, L.K., et al., *Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight.* Nature communications, 2019. 10(1): p. 1893.

149. Chung FF, Herceg Z. *The Promises and Challenges of Toxico-Epigenomics: Environmental Chemicals and Their Impacts on the Epigenome*. Environ Health Perspect. 2020. 128(1):15001.

II. Hypothesis

A variety of external and internal exposures affect human development and disease outcomes, including CC. The latency period between exposure and CC is relatively short, hence, facilitating follow-up from the time of exposure till disease onset. Moreover, the *in utero* period is a particularly vulnerable window to exposure because of the large capacity of changes in cell fate that could occur at this time point, with lifelong consequences. Epigenetic mechanisms are crucial herein because of their heritable nature and driver role in embryogenesis, dictating the fate of the various cell types which otherwise have identical genetic makeups (**Figure 4**). Therefore, **we hypothesized that epigenetic mechanisms, particularly the well-established DNA methylation, underlie biological pathways linking early-life factors to later onset of CC.**

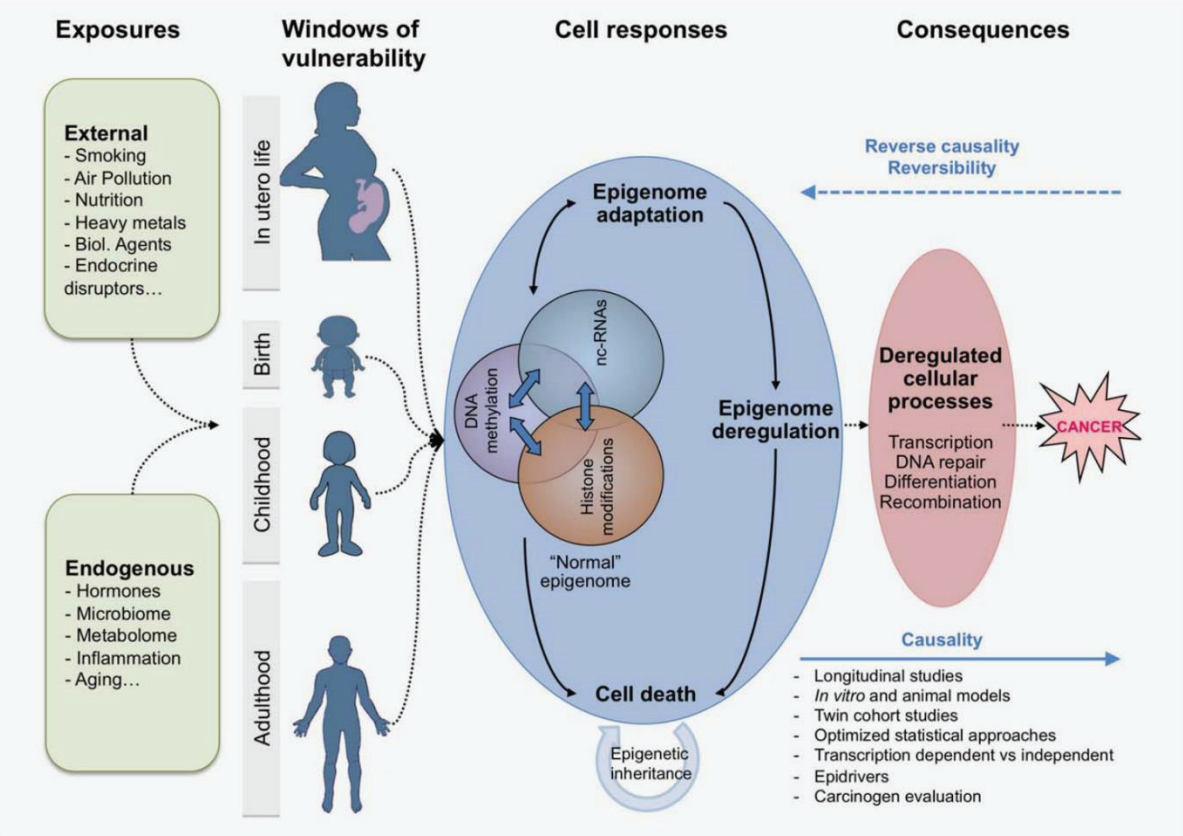


***Figure 4**. Impact of in utero exposures on epigenetic reprogramming and the developmental origin of cancer. Adapted from (52)*

III. Specific aims

Aim 1: Characterize neonatal epigenomic biomarkers of early-life intrinsic factors and "intermediate" phenotypes that predispose to later onset of CC. Detailed exposure modelling and data harmonization was performed across cohorts. Among the early-life factors (E), the thesis prioritized closely related intrinsic factors: BW, gestational age and child sex. The frequent use in studies of BW for gestational age (e.g. when studying different geographical regions) and of BW stratification by child sex (as BW is higher among males) endorse the tight relationship between these three variables. Furthermore, BW is considered a collective proxy of overall *in utero* exposure and is so far one of the few risk factors robustly associated to CC through both retrospective and prospective evidence (see Section I.B.3). Hence, BW represents an "intermediate" phenotype occurring during the latency period between exposure time and cancer onset. An important next step forward would be to study the mechanisms by which these factors associate to CC. Epigenetics (DNA methylation) was the major mechanism investigated due to its driver role during embryogenesis (the major time point herein), heritable nature (hence, long-lasting effects), reversible potential (hence, preventable) and known ability to act as molecular sensors of the environment (52). The DNA methylome was profiled (using Illumina Infinium bead arrays) on neonatal blood samples using a large sample from population-based birth cohorts (with abundant non-diseased subjects). Because blood represents a heterogeneous mixture of cells, confounding due to cell types was adjusted for by using reference-based (139) and reference-free methods (e.g. surrogate variable analysis).

Aim 2: Identify DNA methylome precursors of childhood cancer in relation to early-life intrinsic factors and "intermediate" phenotypes. Among the CC, ALL was focused on because it is the most common subtype of CL and because otherwise including all CC would increase heterogeneity. A dataset based on a relatively homogeneous and largest subtype will likely have higher statistical power. Linking early-life exposures to ALL through epigenetic mechanisms requires sufficient numbers of "exposed" cancer cases, particularly given the large dimension of statistical tests of the human methylome. Therefore, we have designed a hypothesis-driven "three-way modelling" approach (see Sections IV and VI) that enables the identification of cross-talk associations between early-life factors, epigenetic mechanisms and cancer outcome. This approach focused on BW and its closely related gestational age and child sex as intrinsic early-life factors.

IV. Study design

As CC is very rare, international effort is crucial to bring together data and biospecimen from multiple cohorts, which is why IARC plays a lead role in the major CC consortia, namely I4C (140) and CLIC (46). I4C provides a unique and the largest platform of prospective data and biospecimens (taken before CC develops) from different cohorts across the globe (and for which IARC is the International Biospecimen Coordinating Centre). I4C encompasses 445,000 subjects, including 650 nested cases, which will be supplemented with additional cases from neonatal blood spot biobanks from the CLIC that are linked to national cancer registries (223 ALL and 227 controls from California, USA, and 111 ALL and 111 controls from Melbourne, Australia, totaling 334 cases and 337 controls). Though not part of cohort studies, those blood spots capture a molecular snapshot of the epigenome before CC developed, so from a mechanistic perspective, they yield prospective evidence that is comparable to that collected from the I4C cohort studies. Methodology was optimized for methylome-wide analysis on neonatal blood samples (137). Additional controls will be included as well from our population-based (i.e. non cancer-enriched) cohort studies that incorporate omics data, namely, the EXPOsOMICS (141) and the PACE (68) consortia. These abundant (healthy) newborns offer sufficient sample sizes, hence statistical power, for high-coverage methylome profiling and adjustment for major confounders. In addition, this approach yields, based on a healthy set of individuals, biomarkers that are not biased by disease status (known as reverse causality) (52) and, by analysing relatively common “intermediate phenotypes”, statistical power is further increased.

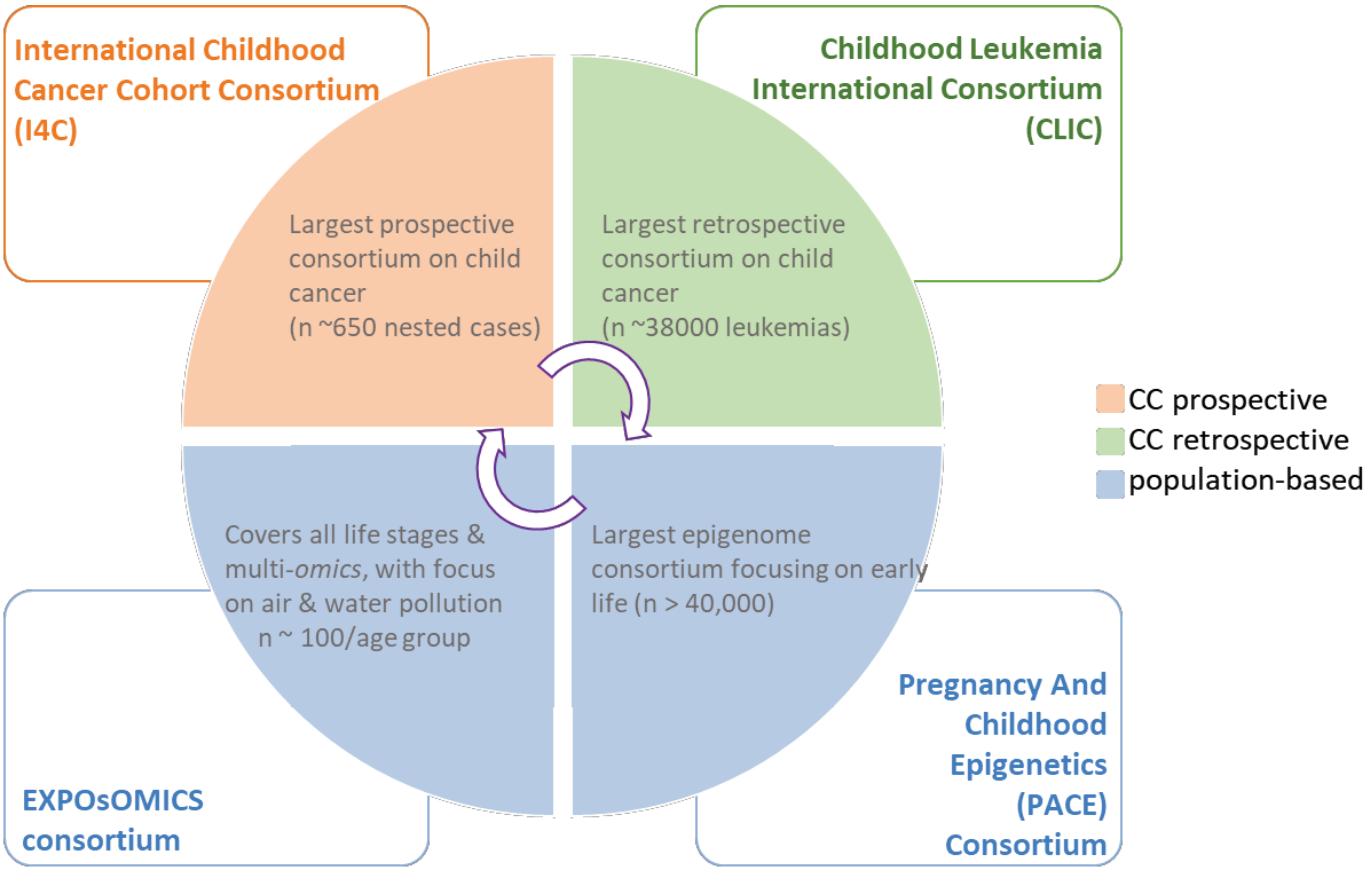


***Figure 5**. Consortia scheme*

The thesis work focused to a large extent on the biostatistics/bioinformatics analysis of the datasets available from the four partner consortia above (**Figure 5**). The datasets comprise questionnaires on lifestyle and exposure factors at birth and perinatally as well as epigenomics (DNA methylome) data. In some cohorts, transcriptomics data was also used to evaluate the functional effect of DNA methylation changes on gene expression as well as other omics data that were integrated altogether to gain further mechanistic insights related to methylation alterations. Based on these datasets, the overarching objective is to decipher molecular pathways linking early-life exposure/lifestyle factors with CC risk.

We tested our hypothesis through our proposed triangulation approach, cross-linking early-life factors (E), epigenetic mechanisms (M) and cancer outcome (C) (**Figure 6**). The orange axis represents the purely epidemiological studies associating early-life factors and CC (and these are being led by the epidemiology partners of our consortium). The purple axis investigates mechanistic precursors in direct association to CC risk, independent of exposure. As this axis represents an agnostic (hypothesis-free) methylome-wide analysis covering ~450,000 CpGs, statistical power represents an important limitation of this approach, which would require additional sample sizes as well as cohorts in order to validate the robustness of findings and minimize potential effects of confounding. The green axis represents associations between

early-life factors and their epigenetic biomarkers, which when identified, can be investigated in relation to CC risk through a hypothesis-driven approach.

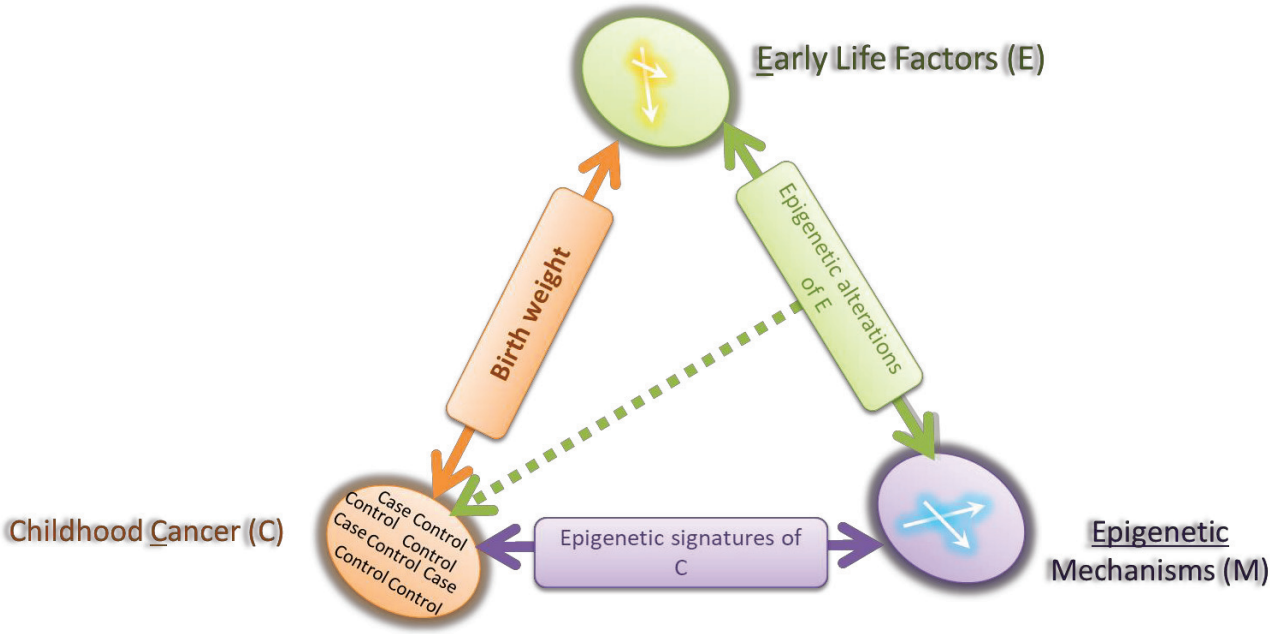


Figure 6. *Triangulation approach, cross-linking E, M and C*

Linking early-life exposures to CC through epigenetic mechanisms requires sufficient numbers of "exposed" cancer cases, particularly given the large dimension of statistical tests of the human methylome. Therefore, we have designed a hypothesis-driven "three-way modelling" approach (see Section VI) that enables cross-talk between the (green) environment–epigenetics axis (such as in PACE/EXPOsOMICS) and the (purple) epigenetics–cancer outcome axis (such as in I4C/CLIC). Briefly, this approach consists of three steps. First, epigenetic biomarkers of early-life factors are screened for in large population-based studies with abundant non-diseased subjects (Aim 1). Second, the specific significant biomarkers of exposure and intermediate phenotypes obtained in the first step are analysed in the (smaller) subset of samples that are enriched for nested cancers (herein, statistical power is maintained due to a targeted low-coverage profiling). Third, statistical modelling such as mediation analysis is used to investigate whether the identified biomarkers mediate the effect of exposure on the cancer outcomes. This approach enables the potential identification of underlying molecular mechanisms linking risk factors to later onset of CC.

V. Methodology

This thesis is part of a broader initiative led by the IARC Epigenetics Group involving large interdisciplinary projects. The choice of appropriate methods in such study settings with data generated using modern high-throughput techniques represents a challenging question. We describe below statistical and bioinformatics approaches developed and applied during the thesis, including epigenomics analysis pipeline, multi-omics data integration and cancer driver prediction tools.

A. Analysis pipeline for methylation data

The key focus was analyzing epigenetic data generated (partially in-house) using Illumina Infinium® HumanMethylation450 BeadChip array, which allows the interrogation of approximately 450,000 sites spanning all Reference Sequence (RefSeq) (142) genes and with proven reproducibility and practicability for studying DNA methylation in large cohort samples (143). However, methylation measures quantified by microarray techniques can be affected by systematic variation due to the technical processing. Thus, we describe below key steps from the in-house optimized pipeline for preprocessing, quality control and batch correction of methylome data.

1. Preprocessing of data

The majority of the results presented in this manuscript were obtained from the analysis of DNA methylation data using the Illumina Infinium® HumanMethylation450 BeadChip assay. Bisulfite conversion was performed using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, CA). From the 500 ng of bisulfite-converted DNA per sample, 250ng was used for hybridization on the 450K array. Each array consisted of 96 samples distributed equally among 8 chips. The study design was established such that batch effects (e.g. sample position and intra- and inter-variability in arrays and chips) did not completely confound with biological covariates. Specifically, each chip included proportional amounts of samples representing different confounder factors (e.g. case-control status, sex, etc.). The bias related to sample position on the chip was also demonstrated in our methodological paper for removing unwanted sources of variation (144). Raw fluorescence intensities (.idat files) were retrieved and preprocessed using the minfi R package (145). To calculate the methylation level of each CpG as the beta-value, we used the following formula:

$$\beta = \frac{\text{intensity of the methylated allele (M)}}{\text{intensity of the unmethylated allele (U)} + \text{intensity of the methylated allele (M)} + 100}$$

β-values, therefore, represent the proportion of methylation at each CpG site (0 = completely unmethylated, 1 = completely methylated). Missing values in methylation data were imputed when their proportion was smaller than 5% per sample or per CpG and removed otherwise. Using raw methylation values, we were able to estimate the sex of the samples. Estimation of child sex is done by using the median values of measurements on the X and Y chromosomes, respectively, and compared with a cut-off. We then compared the estimated sex with sex status from questionnaires. Sex mismatches were consistently excluded from further analysis.

2. Quality control

One of the quality control (QC) steps involve use of getQC function to identify bad quality sample. Samples with a lower QC than the threshold are flagged as bad quality and removed from further analysis. To further explore the quality of the samples, it is useful to look at the methylation values densities of the samples. Indeed, a minor number of cases can pass the above-mentioned threshold but still display an unusual methylation distribution density of the 450K CpG probes. Visualizing box plots of both methylated (M) and unmethylated (U) channels allows us to verify the presence or absence of biases among samples. Furthermore, multi-dimensional scaling (MDS) plot showing a 2-D projection of euclidean distances can pinpoint cluster of samples, including potential outliers.

Although different normalization procedures are available for the Illumina Infinium® HumanMethylation450 platform, we used Functional normalization (Funnorm, minfi R package) that was shown to perform equally good or outperform existing normalization methods (146). It removes unwanted variation by regressing out variability explained by the control probes present on the array.

The efficiency of the normalization can be depicted by replotting the methylation distribution of 450K CpG probes. This is followed by filtering of cross-reactive probes (that target repetitive sequences or co-hybridize to alternate sequences, thus, generating spurious signals), and sometimes of CpGs on sex chromosomes or of single-nucleotide polymorphism (SNP)-related CpG probes. Finally, we use surrogate variable analysis (SVA) (147) for batch correction, a choice validated by the findings of our benchmarking (144). SVA is also used as a reference-free method to adjust for differences in white blood cell (WBC) composition.

WBC can be also predicted based on regression calibration algorithms (121) from 450K microarrays using reference data pertinent to peripheral (148) or cord blood (149), which are the tissue types investigated in our work. For example, deconvoluting methylation data from blood will result in relative proportions of CD4+ and CD8+ T-cells, natural killer cells, monocytes, granulocytes, and B-cells in each sample, and these proportions can be adjusted for in subsequent analyses.

3. Statistics

A first step into the statistical investigation of the data is to proceed to CpG site-by-site regression analysis. To reduce the number of false positive results, we control for

multiple testing using Benjamini Hochberg procedure (150) or more stringent Bonferroni approach (151). In order to minimize further the false positive rate and in case of several participant cohorts, a meta-analysis is conducted by one centre and re-assessed by a second « shadow » meta-analysis performed by independent group. We used METAL software (152) to perform a fixed (or random) effect with inverse variance based meta-analysis. Besides performing site-by-site analysis, we also used a dimension reduction approach (DMRcate) (153) to reduce the 450K individual sites into clusters of genetically proximal and correlated CpGs to enhance statistical power and aid biological inferences (as single CpG sites often have more subtle functional relevance than CpG clusters) , as per our previous work (154,155).

Additionally, in some studies, we integrated DNA methylation data with other types of omics when available. Separate analyses of each data source capture important features that are specific to each source; whereas, joint analysis offers the opportunity to highlight the shared variation across several omics, which is particularly important especially in multi-layered diseases like cancer. The demand for such integrative methods motivates a dynamic area of statistics and bioinformatics, and we have benchmarked some major methods in this regards (156).

B. Relevant publications

The following papers describe three major methodology investigations performed as part of the PhD thesis. The first paper aims to identify and correct technical biases present in 450K methylation data. As epigenetics was the primary focus of the analyses and main source of findings, this step was of paramount importance. Indeed, before trying to biologically interpret the epigenetic findings, we need to be confident that upstream analyses were of highest quality. The second co-led paper comprises work started during the master internship and continued during the thesis. It focuses on extracting valuable information across several omic layers. Indeed, recent technical advances facilitated the collection of large-scale omics data from the same biological samples. As will be seen, the studies presented in the results chapter required skills in extracting information from epigenomics but also other types of omics. The third paper is a pan-cancer investigation using genomic and transcriptomic data for the assessment of epigenetic cancer driver genes. This study builds on a battery of novel biostatistic and bioinformatic tools in order to perform integrated analysis of the two omics. It required, inter alia, knowledge acquired in the first two papers and generated cancer driver prediction tools that can be applied in future studies on the epigenetic markers identified in this work in order to investigate whether such markers play a driver role in pediatric cancers (in addition to serving as early biomarkers detectable before disease onset).

1) Perrier F, **Novoloaca A**, Ambatipudi S, Baglietto L, Ghantous A, Perduca V, Barrdahl M, Harlid S, Ong KK, Cardona A, Polidoro S, Nøst TH, Overvad K, Omichessan H, Dollé M, Bamia C, Huerta JM, Vineis P, Herceg Z, Romieu I,

Ferrari P. Identifying and correcting epigenetics measurements for systematic sources of variation. Clin Epigenetics. 2018 Mar 21;10:38.

2) Chauvel C*, **Novoloaca A***, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 2019 Feb 14. pii: bbz015.
*Co-first authorship.

3) Halaburkova A, Cahais V, **Novoloaca A**, Khoueiry R, Ghantous A, Herceg Z. Pan-cancer genome and transcriptome analysis and orthogonal experimental assessment of epigenetic driver genes. In review in Genome Research.

Perrier et al. Clinical Epigenetics (2018) 10:38
https://doi.org/10.1186/s13148-018-0471-6

Clinical Epigenetics

METHODOLOGY Open Access

Identifying and correcting epigenetics measurements for systematic sources of variation

Flavie Perrier[1], Alexei Novoloaca[2], Srikant Ambatipudi[2], Laura Baglietto[3], Akram Ghantous[2], Vittorio Perduca[4], Myrto Barrdahl[5], Sophia Harlid[6], Ken K. Ong[7], Alexia Cardona[7], Silvia Polidoro[8], Therese Haugdahl Nøst[9], Kim Overvad[10,11], Hanane Omichessan[12,13], Martijn Dollé[14], Christina Bamia[15,16], José Marìa Huerta[17,18], Paolo Vineis[19], Zdenko Herceg[2], Isabelle Romieu[20] and Pietro Ferrari[1*]

Abstract

Background: Methylation measures quantified by microarray techniques can be affected by systematic variation due to the technical processing of samples, which may compromise the accuracy of the measurement process and contribute to bias the estimate of the association under investigation. The quantification of the contribution of the systematic source of variation is challenging in datasets characterized by hundreds of thousands of features. In this study, we introduce a method previously developed for the analysis of metabolomics data to evaluate the performance of existing normalizing techniques to correct for unwanted variation. Illumina Infinium HumanMethylation450K was used to acquire methylation levels in over 421,000 CpG sites for 902 study participants of a case-control study on breast cancer nested within the EPIC cohort. The principal component partial R-square (PC-PR2) analysis was used to identify and quantify the variability attributable to potential systematic sources of variation. Three correcting techniques, namely ComBat, surrogate variables analysis (SVA) and a linear regression model to compute residuals were applied. The impact of each correcting method on the association between smoking status and DNA methylation levels was evaluated, and results were compared with findings from a large meta-analysis.

Results: A sizeable proportion of systematic variability due to variables expressing 'batch' and 'sample position' within 'chip' was identified, with values of the partial R^2 statistics equal to 9.5 and 11.4% of total variation, respectively. After application of ComBat or the residuals' methods, the contribution was 1.3 and 0.2%, respectively. The SVA technique resulted in a reduced variability due to 'batch' (1.3%) and 'sample position' (0.6%), and in a diminished variability attributable to 'chip' within a batch (0.9%). After ComBat or the residuals' corrections, a larger number of significant sites ($k = 600$ and $k = 427$, respectively) were associated to smoking status than the SVA correction ($k = 96$).

Conclusions: The three correction methods removed systematic variation in DNA methylation data, as assessed by the PC-PR2, which lent itself as a useful tool to explore variability in large dimension data. SVA produced more conservative findings than ComBat in the association between smoking and DNA methylation.

Keywords: Epigenetics, PC-PR2, Normalization, Methylation, Smoking status

* Correspondence: ferrarip@iarc.fr
[1]Nutritional Methodology and Biostatistics Group, International Agency for Research on Cancer (IARC), World Health Organization, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France
Full list of author information is available at the end of the article

Background

Epigenetics aims at investigating changes in gene activity not attributable to changes in the DNA sequence [1]. An increasing number of studies analysed epigenetics in relation to modifiable environmental exposures of epidemiologic interest, such as smoking [2–4], alcohol consumption [5], maternal plasma folate [6] and other vitamin involved in the one carbon metabolism pathway [7], as well as the role of epigenetic profiles on the risk of developing chronic diseases, including cancer [8]. DNA methylation is a mechanism of epigenetic regulation that involves the addition of methyl groups (–CH3) to the cytosine of a cytosine-guanine DNA sequence. DNA methylation level at one CpG site is frequently expressed as the percentage of cells that are methylated at that specific site. The Illumina Infinium HumanMethylation450K BeadChip (HM450K) quantifies DNA methylation at more than 450,000 interrogated CpG sites, expressing methylation level as the ratio of the methylated probe intensity to the overall intensity, which is the sum of the methylated and unmethylated probe intensities [9].

Methylation levels are influenced by many factors including aging [10] and environmental exposure [11, 12], but might also be affected by systematic variation due to the processing of the biospecimens, e.g. variability attributed to batch (a sub-group of samples processed at the same time, 96 samples per batch in the HM450K), chip position within batches (8 chips per batch in the HM450K) and the position of the samples within the chip [13]. Methods of correcting for the sources of methylation variability include ComBat, based on an empirical Bayes method [14] and the surrogate variables analysis (SVA) [15, 16]. An alternative method consists in the computation of residuals from a beta regression, where methylation levels were regressed on the major sources of methylation variability.

The large dimension of new generation methylation arrays makes it difficult to quantify the amount of variability attributable to systematic sources of variation. The principal component partial R-square (PC-PR2) method was developed to quantify the contribution of sources of variation defined a priori in large dimensional data [17].

Smoking exposure has been analysed in many studies [2–4], which offers a large comparative pool of results. Smoking has also been shown to have a major impact on the epigenome and hence provides a large number of significant CpGs to analyse. For these reasons, in this work, we have chosen to evaluate the performance of ComBat, SVA and the residuals' method to correct for potential systematic variability in methylation measurements, in the association between smoking and DNA methylation levels from DNA samples of subjects of a nested case-control study on breast cancer conducted within the European Prospective Investigation into Cancer and nutrition (EPIC) study. The PC-PR2 method was used to quantify the extent of total epigenetics variability before and after applying each correcting method.

Methods

Study population

The EPIC study [18, 19] is a multicentre study that recruited over 521,000 study participants, between 1992 and 2000 in 23 regional or national centres in 10 European countries (Denmark, France, Germany, Greece, Italy, Netherlands, Norway, Spain, Sweden and the UK). Among the 367,903 women recruited in EPIC, we excluded 19,583 participants with prevalent cancers at recruitment (except non-melanoma skin cancer) and 2892 women that were lost during follow-up. Malignant primary breast cancer (BC) occurred for 10,713 of them from 1992 to 2010. A nested case-control study was designed among women who completed dietary and lifestyle questionnaires and provided blood samples at recruitment (baseline), which included 3858 invasive BC cases. Each case was matched to a randomly selected control among cancer-free women by recruitment centre and the following baseline variables: age, menopausal status, fasting status, current use of oral contraceptive pill or hormone replacement therapy and time of blood collection [20].

Genome-wide DNA profiling assessment

Genome-wide DNA-methylation profiles in buffy coat samples was quantified using the Illumina Infinium HumanMethylation450K (HM450K) BeadChip assay [9] in 960 biospecimens of women included in the BC nested case-control study [21]. The 480 cases were selected based on estrogen receptor status and by selecting equal proportions of subjects with above or below median level of dietary folate. Matched controls were the same than those selected for the whole study. A total of 20 biospecimens with replicates were used to compare technical inter- and intra-assay batch effects and then excluded from the main analysis. We also excluded 19 matched pairs where at least one of the two samples had a low-quality bisulfite conversion efficiency (intensity signal < 4000) or which did not pass all the Illumina GenomeStudio quality control steps, which were based on built-in control probes for staining, hybridization, extension, and specificity [22]. A total of 451 completed matched pairs ($n = 902$) were retained for the main statistical analyses. In any given sample, probes with detection p value higher than 0.05 were assigned 'missing' status. After the exclusion of 14,548 cross-reactive probes, 47,963 probes overlapping known SNPs with minor allele frequency (MAF) of ≥ 5% in the overall population (European ancestry) [23] and 1483 low-

quality probes (missing in more than 5% of the samples), 421,583 probes were included in the statistical analyses.

For each probe, β value was calculated as the ratio of methylated intensity and the overall intensity, defined as the sum of methylated and unmethylated intensities. The following preliminary adjustment steps were applied to the β values: (i) color bias normalization using smooth quantile normalization to correct for the two color channels; (ii) quantile normalization [24]; (iii) type I and type II bias correction using the beta-mixture quantile normalization (BMIQ) [25]. Then, M values, defined as $M_{\text{values}} = \log_2(\frac{\beta_{\text{values}}}{1-\beta_{\text{values}}})$, were computed [26]. In this work, the β and M values obtained after the preliminary normalization steps were referred to as the raw β and M values.

The amount of white blood cell counts (T cells ($CD8^+T$ and $CD4^+T$), natural killer (NK) cells, B cells, monocytes and granulocytes) was quantified using Houseman's estimation method [27]. The percentage of granulocytes was not included in this analysis as it is collinear with the five other white blood cell counts: the total of the percentages of the six leukocyte subtype counts is 1.

For the DNA methylation measurements with the HM450K BeadChip, samples were aliquoted into 10 batches; each batch was made of 8 chips, and each chip contained 12 samples (located in 2 columns of 6 rows). Chip position represented the position of the chips within a batch, as illustrated in Fig. 1a, and sample position represented the position of the samples within a chip, as in Fig. 1b.

Lifestyle exposures

Data on lifestyle exposures were collected at recruitment through country- or centre-specific dietary and lifestyle questionnaires [18]. Smoking status was categorized into ever (former/current) and never smokers and was not associated to any of the technical covariates.

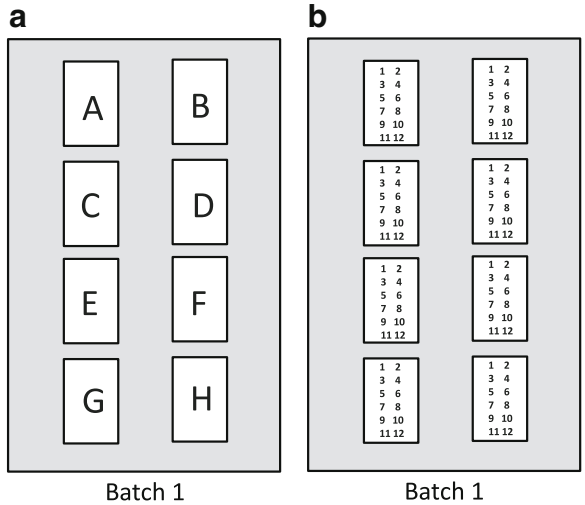


Fig. 1 Description of laboratory variables. **a** Position of chips within batches, each batch was made of 8 chips. **b** Sample position within chips, each chip contains 12 samples

Statistical analyses

In order to inspect the variability of DNA methylation levels, we first visually inspected, via box plots, global DNA methylation levels by batch, chip and sample positions. The principal component partial R-square (PC-PR2) method was used to quantify the contribution of laboratory factors and other characteristics of the samples to the between-sample variability observed [17]. First, principal component analysis (PCA) was carried out, by the PC-PR2, on the matrix X of epigenetics data of dimension $n \times p$ ($n = 902$: number of study samples and $p = 421{,}583$: number of probes). In PCA, eigenvalues and eigenvectors are usually obtained from the matrix $X'X$ of dimension $p \times p$. In this case, and in general with *-omics* data, p is very large ($p \gg n$), and the decomposition of $X'X$ can be cumbersome. A particularly appealing procedure consists in extracting eigenvalues and eigenvectors from the matrix XX', of dimension $n \times n$ [28], which is way easier to handle, being n much smaller than p. Once eigenvalues were extracted, the q first components explained an amount of total variability in X greater than a given threshold, i.e. 80% in this study. Then, each of the q first PCA score components was, in turn, linearly regressed on a list of independent covariates (Z), comprising of laboratory factors and characteristics of the samples. Values of the partial R^2 statistics were assessed for each Z covariate, separately in each component-specific model [29]. An overall partial R^2 was computed for each Z covariate with a weighted average of their component-specific partial R^2 using the corresponding q eigenvalues as weights, conditional to all other covariates in the model. The covariates that we have entered into the regression include batch, chip position, row sample position, recruitment centre, proportions of leukocyte subtypes ($CD8^+T$, $CD4^+T$, NK, B cells and monocytes), alcohol consumption (g/day), age (year), BMI (kg/m^2), menopausal status (post- vs. pre-menopause), smoking (ever vs. never smokers), BC status (case or control) and dietary folate intake (μg/day).

Removing unwanted variation

To remove the two most important sources of variation identified with the PC-PR2 from DNA methylation levels, three different correcting techniques were applied to raw β and M values: residuals, ComBat and SVA. The ComBat method [14] is a procedure based on an empirical Bayes approach that can correct only for one covariate at the time. Given the presence of multiple sources of variation, we have applied two parametric ComBat in multiple sequential steps: ComBat was first applied to remove batch variability, and then a second ComBat step

was run to remove variability due to row sample position. Methylation β values that after the application of ComBat were lower than 0 or larger than 1 were set to 0 and 1 respectively. The surrogate variables analysis (SVA) is a method developed to remove pre-identified sources of variability but also non-known sources of variability, i.e. variability which is not specified in the SVA model, using surrogate variables [15, 16]. Once surrogate variables were assessed by SVA, residuals from a regression modeling methylation level according to the surrogate variables were computed to remove the unwanted variation.

As the β values are continuous in the [0,1] interval, the calculation of the residuals for the residuals' method and SVA method were based on beta regression. To be comparable to the ComBat and raw (i.e. uncorrected) data, residuals computed with the residuals' and the SVA methods needed to be rescaled as follows:

$$res_{\text{scaled},j} = \frac{\text{res}_{\text{raw},j} - \min\left(\text{res}_{\text{raw},j}\right)}{\max\left(\text{res}_{\text{raw},j}\right) - \min\left(\text{res}_{\text{raw},j}\right)} \left(\max\left(\text{raw}_j\right) - \min\left(\text{raw}_j\right) \right) + \max\left(\text{raw}_j\right)$$

where $j = 1\ldots421{,}583$, raw_j represents the raw β values measured in site j and $\text{res}_{\text{raw},\ j}$ the residuals computed for site j before transformation.

In order to check the efficacy of the three correcting techniques, a second PC-PR2 analysis was used to quantify the contribution of each laboratory factor to total variability, after each of the normalization methods.

Same approach was used for M values using a linear regression instead of beta regression to compute residuals from the residuals' and the SVA methods.

In order to compare sample individual values before and after correction, raw and corrected β and M values of the probe cg00000029 were visually inspected. In this site, in addition to the three tested methods, a second residuals' method was also computed using random effects instead of fixed affects to remove unwanted variation, from a beta or linear mixed regression, respectively for β and M values.

CpG site-specific models

The association between smoking status and each of the 421,583 CpG sites was carried out before and after application of each normalization method. Beta regression models were used for β values and linear regression models for M values, with adjustment for chip position, recruitment centre, percentages of five leukocyte subtypes, age at recruitment, menopausal status and BC status. The standard adjustment models, i.e. models using the raw methylation values, were also adjusted for batch and row sample position. In order to compare the epigenome-wide distribution of p values with the expected null distribution of p values, the inflation factor λ was computed and the quantile-quantile (QQ) plots were generated. The inflation factor was defined as the ratio of the median of the observed $\log_{10}$ transformed p values and the median of the expected $\log_{10}$ transformed p values. False discovery rate (FDR) was used to control for multiple testing. In order to compare the performance of the different correction methods with a nominal reference, the list of k significant CpG sites (q values < 0.05) associated with smoking was compared to the results of a large meta-analysis carried out in the CHARGE consortium, a recent large meta-analysis on the link between the epigenetic signature of cigarette smoking that pooled data from 16 studies, and included about 16,000 individuals [4]. In CHARGE, smoking status was statistically significantly associated with DNA methylation level (β values) in 18,760 sites, after FDR correction of p values.

In order to compare the performance of the correction methods, the relative sensitivity and specificity of each correcting method were computed. We considered the CpG sites significantly associated to smoking in the CHARGE consortium as the true positives, i.e. an arbitrary gold standard, given that this is a well-powered reference study and the largest to date.

Preprocessing steps and statistical analysis were carried out using the R software (https://www.r-project.org/) and Bioconductor packages [30], including 'lumi' and 'wateRmelon' for the adjustment step, 'sva' [31] for ComBat and SVA corrections, and 'betareg' for beta regression models. The PC-PR2 method was computed using the R code available in Fages et al.'s supplementary material [17].

Results

DNA measurements of the first and the last batches were conducted roughly 3 months apart. DNA measurement of two consecutive batches varied from 3 to 14 days. Box plots of global methylation (i.e. mean of methylation levels in all the CpG sites) showed a random variation of global methylation levels between batches, as reported in Fig. 2a for β values. Global methylation between chip positions did not present large variation (Fig. 2b). Sample position within the chip systematically influenced global methylation, with levels by rows, showing a progressive constant increase in methylation, a feature not observed by column, as displayed in Fig. 2c. The impact of row sample position on global methylation was even stronger when batches were evaluated separately (Fig. 2d). Global methylation computed with M values gave similar results (Additional file 1: Figure S1).

Tables 1 and 2 show the results of PC-PR2 to quantify the amount of total variability of DNA methylation explained respectively by laboratory factors and characteristics of the samples (recruitment centre, the five

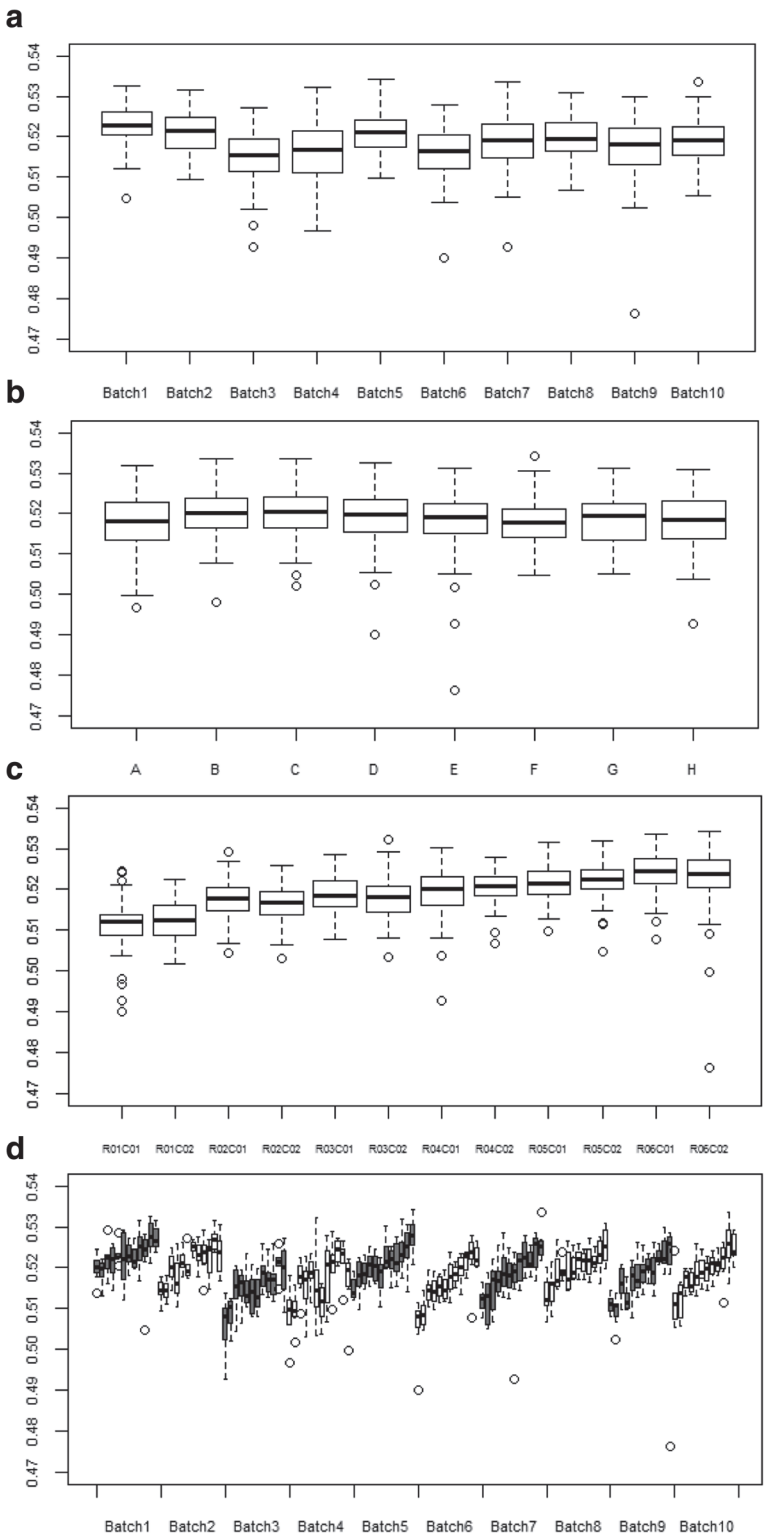


Fig. 2 Box plots of global methylation (β values) according to laboratory factors. **a** Batch. **b** Chip position within batches. **c** Sample position within chips. **d** Batches and sample position within chips

Table 1 Values of weighted partial R^2 (%) from PC-PR2 analysis indicating the proportion of variability of methylation levels, before and after normalization step, explained by a specific set of laboratory factors

Values	Methods[a]	Row sample position	Batch	Chip position	Total[b]
β values	Raw	11.4	9.5	6.5	30.4
	Residuals	0.2	1.3	5.9	17.9
	ComBat	0.2	1.3	6.0	17.1
	SVA	0.6	1.3	0.9	6.5
M values	Raw	12.3	9.7	6.8	30.7
	Residuals	0.2	1.2	5.8	16.5
	ComBat	0.2	1.3	6.2	17.0
	SVA	0.4	0.7	0.8	5.3

[a]Residuals, COMBAT and SVA methods used to correct effect due to batch and row sample position (within the chips)
[b]Total variability explained by laboratory factors and characteristics of the samples (recruitment centre, the five percentages of leukocyte subtypes, alcohol consumption, age and BMI, menopausal status, smoking, BC status and dietary folate)

percentages of leukocyte subtypes, alcohol intake, age, BMI, menopausal status, smoking, breast cancer status and diet folate intake), for raw β and *M* values. Findings were similar for raw β and *M* values; the largest contribution to the overall variability came from row sample position and batch explaining, respectively, 11.4 and 9.5% (β values), and 12.3 and 9.7% (*M* values) of overall methylation variation. Chip position contributed to 6.5 and 6.8%, for raw β and *M* values respectively. The percentages of leukocyte subtypes and centre explained most of the variation of DNA methylation due to sample characteristics for raw β and *M* values. Each of the remaining tested other sample characteristics explained less than 0.5% of total variation.

Table 2 Values of weighted partial R^2 (%) from PC-PR2 analysis indicating the proportion of variability of raw methylation levels explained by a specific set of covariates

Characteristics of samples	β values	*M* values
Recruitment centre	3.0	2.9
Percentages of leukocyte subtypes		
CD4T	3.2	3.2
CD8T	3.7	3.1
Natural killers	5.2	4.7
B cells	1.7	1.1
Monocytes	0.4	0.4
Alcohol intake at recruitment	0.2	0.1
Age at recruitment	0.4	0.4
BMI at recruitment	0.1	0.1
Menopausal status	0.2	0.2
Smoking status	0.1	0.2
Breast cancer status	0.1	0.1
Dietary folate	0.1	0.1

Removing unwanted variation

All the three correcting methods decreased the contribution of row position and batch to similar neglectable levels, whereas only SVA appeared to reduce the contribution to variability due to chip position (Table 1). The amount of variability explained by laboratory factors and sample characteristics for raw β values decreased from 30.4 to 17.9% and 17.1% using, respectively, the residuals' method and ComBat, and to 6.5% after SVA. The PC-PR2 approach applied on *M* values estimated values of partial R^2 for laboratory factors and sample characteristics similar to those of β values.

Corrected methylation values of the probe cg00000029 were very similar using ComBat or the residuals' methods for β values and *M* values (Fig. 3). SVA corrected values were the corrected values most different from the raw values. Using the residuals' method with fixed or random effects for batch and row sample position gave similar results.

CpG site-specific models

The frequency *k* of sites associated with smoking status is shown in Table 3, consistently for β and *M* values. For β values adjusted by batch and row sample position (standard adjustment), smoking status was significantly associated to methylation levels in 444 sites. The number of CpG sites significantly associated with smoking status was equal to 427 for the residuals' method, 600 for ComBat and 96 for SVA after correction. According to the inflation factors and QQ plots, there was no evidence of inflation for any methods (Additional file 2: Figure S2).

These frequencies were compared to the list of 18,760 sites identified in the CHARGE meta-analysis (Joehanes et al. [4]). A total of 77 sites overlapped across the standard adjustment and the three correcting methods in this study and the sites identified in the consortium, as shown in the Venn diagram for β values in Fig. 4a. In addition to these sites, the standard adjustment, the residuals' method and the ComBat method shared a list of 249 significant sites with CHARGE. The ComBat method resulted in the largest frequency of sites overlapping with results in CHARGE ($k = 411$), but also in the largest percentage of sites not observed in CHARGE (31%). In contrast, SVA identified the lowest number of significant sites ($k = 96$) but the vast majority of them (92%) were also identified in CHARGE.

As for *M* values, 322 sites were associated to smoking using the standard adjustment, $k = 332$ after the residuals' method, $k = 387$ using ComBat, $k = 144$ after SVA correction. A total of 111 sites overlapped all the methods and CHARGE, as shown in Fig. 4b. SVA was

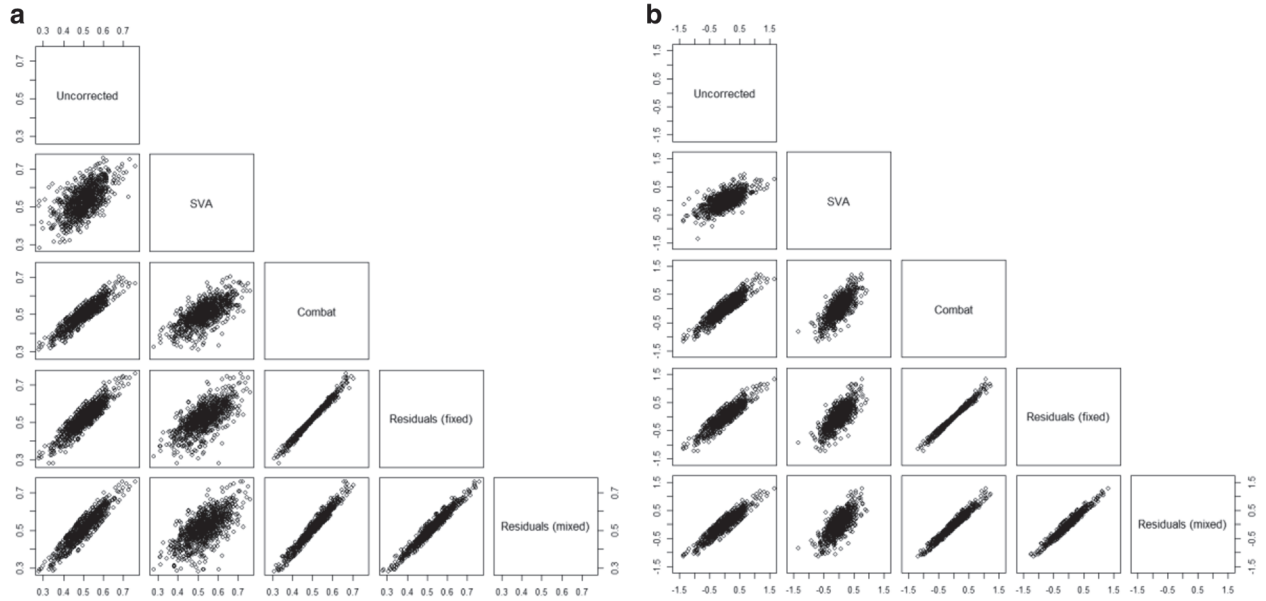


Fig. 3 DNA methylation levels of the CpG site cg00000029 before and after normalization step. **a** β values. **b** *M* values

the method leading to the lowest number of significant sites, but also to the largest percentage of sites also identified by CHARGE (93%). This percentage ranged between 85 and 90% for all the other methods. According to the inflation factors and QQ plots, there was no evidence of inflation for any methods for *M* values (Additional file 3: Figure S3). SVA showed the least inflation in both β values and *M* values.

Sensitivity was similar for the standard adjustment, the residuals' method and the ComBat method with a value about 0.020 for β values and over 0.015 for *M* values (Table 3). SVA sensitivity was four times less for β values and twice less for *M* values. SVA was the most specific method with 1-specificity equals to 0.2×10^{-4} for β values and *M* values whereas ComBat was the least specific with 1-specificity equals to 4.7×10^{-4} and 1.3×10^{-4} for β values and *M* values, respectively.

Table 3 CpG site-specific regression models before and after normalization step

Values	Methods	Significant sites[b]	CHARGE[c]	Sensitivity	1-Specificity
β values	Standard adjustment[a]	444	357 (80%)	1.9×10^{-2}	2.2×10^{-4}
	Residuals	427	365 (85%)	1.9×10^{-2}	1.5×10^{-4}
	ComBat	600	411 (69%)	2.2×10^{-2}	4.7×10^{-4}
	SVA	96	89 (92%)	0.5×10^{-2}	0.2×10^{-4}
M values	Standard adjustment[a]	322	274 (85%)	1.5×10^{-2}	1.2×10^{-4}
	Residuals	332	299 (90%)	1.6×10^{-2}	0.8×10^{-4}
	ComBat	387	335 (87%)	1.8×10^{-2}	1.3×10^{-4}
	SVA	144	134 (93%)	0.7×10^{-2}	0.2×10^{-4}

Models are adjusted for chip position, recruitment centre, the five percentages of leukocyte subtypes and age at recruitment, menopausal status and BC status
[a]Also adjusted for batch and sample position
[b]Number of significant sites for smoking status after *p* values FDR correction
[c]Number (and percentage) of significant sites identified by the CHARGE meta-analysis

Discussion

Batch effects on DNA methylation measurements have already been documented [13]. Various correcting methods have been recently used, including standard adjustment [3], ComBat [6] and SVA [2]. Our findings suggested that batch was not the only source of variation in the DNA methylation data from our EPIC study, as the position of the sample within the chip and, to a lesser extent, chips within batches, also contributed to total variability. Noteworthy, while variation by batch was essentially random, the position of the sample within the chip contributed systematic variation, with methylation levels progressively increasing by row, but not by column. This might be due to the washing step which is done row by row in each chip during the measurement of DNA methylation using HM450K. Eventually, batch and row sample positions explained cumulatively more than 20% of the methylation levels and were the most important sources of variation. Further replications are needed in others dataset from other labs to validate our findings.

PC-PR2 is a powerful method to identify and quantify random and systematic sources of variation in large-scale datasets. Here, the method, initially developed for metabolomics data [17], was successfully applied to epigenetics data, a challenging set characterized by hundreds of thousands of features, and can easily be extendable to other *-omics* data. It is based on the

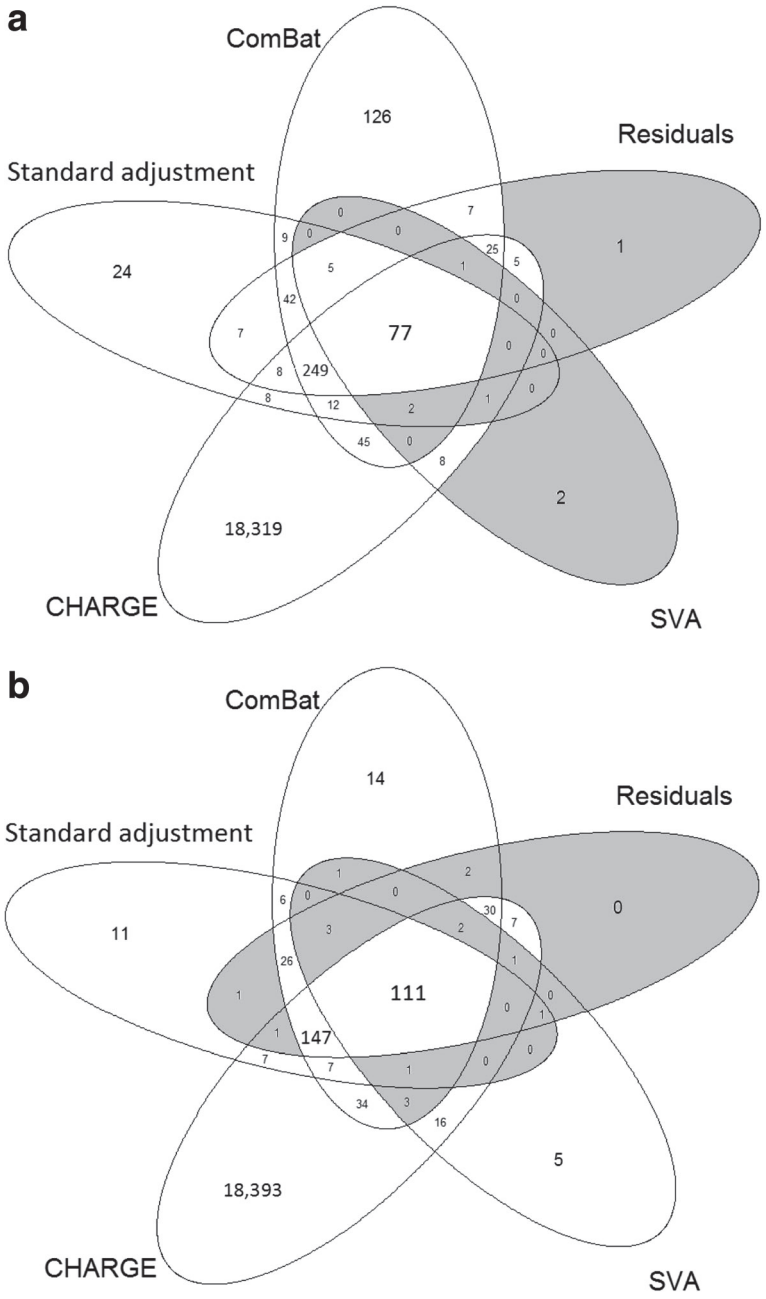


Fig. 4 Venn diagram of significantly identified CpG sites for smoking status using each correcting methods and CHARGE. **a** β values. **b** *M* values. *p* values were corrected for multiple testing with FDR

combination of a principal component analysis (PCA) and the concept of partial R^2 in multivariable linear regression. PC-PR2 quantifies the contribution of variability of continuous and/or categorical covariates to total variability in the outcome data, and in general offers high level of flexibility to capture specific features such as, say, non-linear effects and longitudinal data. A particularly appealing feature is the possibility of performing PCA by decomposing the matrix XX' of dimension $n \times n$ rather than $X'X$ of dimension $p \times p$ that would be virtually untreatable in the *-omics* domain. The PC-PR2 can also be extended to the Infinium MethylationEPIC BeadChip (850K), which is the updated version of HM450K.

Identifying unwanted sources of variation in epigenetics data is a crucial step prior to statistical analysis. Each of the three tested methods succeeded to correct DNA methylation levels for the pre-specified sources of variability.

Percentages of variability due to batch and row sample position diminished to marginal levels after the use of the three methods. Other unknown or unmeasured experimental conditions are also likely to modify DNA methylation measurements, such as differences in sample handling and preparation and the room temperature during sample processing. Overall, the procedures for sample treatment are way more challenging to control, possibly because detailed information on each sample are not always documented, and it is rather assumed that these are relatively homogeneous across recruitment centres. Statistical adjustment for centre is a standard practice in the analysis of epigenetics data and of any laboratory measurements. In this respect, SVA turned out to provide a correction on top of the pre-specified sources of variability through the estimation of surrogate variables possibly influencing overall variability. It was remarkable that the variability attributed to chip position, whose partial R^2 values was 6.5% in the raw data, decreased to 0.9% after SVA, even if chip position was not included in the list of covariates of which we want to remove the variability, specified in the SVA model. Indeed, the surrogate variables, computed by a PCA step in the SVA algorithm, capture the variability in the methylation data which is not already explained by the a priori list of covariates (batch and row sample position). A challenge of DNA methylation data is the presence of outliers that can generate spurious associations. Techniques have been introduced to filter out outliers through preliminary quality control checks globally on all CpG sites [32]. This was achieved through the Illumina GenomeStudio quality in the present study [22]. Nevertheless, outlier values passed the GenomeStudio quality control screening and were detected after applying the residuals or SVA methods. On the contrary, ComBat is based on an empirical Bayesian procedure with an additive and a multiplicative component, the latter contributing to shrink all observations, including outliers [14]. This makes ComBat an attractive solution to control outlier values in large-dimension data. Another interesting feature is that ComBat preserved the observed variability of methylation data in the $[0, 1]$ interval for β values, unlike the residuals' and SVA methods, for which the corrected values could fall outside the $[0, 1]$ range.

The performance of the various correction methods was evaluated in this study through the comparison with results of association between smoking and methylation from the CHARGE consortium, one of the largest studies available to date. This could be a debatable choice but allowed a reference group to be established to compute relative sensitivity and specificity of each normalizing method. The low sensitivity across all methods in our analysis might be explained by the lack of power due to the sample size: over 16,000 samples were included in CHARGE against 902 in our study. Some different characteristics of our population and the one of the CHARGE consortium might also explain the difference in terms of significant sites. For example, only women are included in our analysis and half of them developed latter a breast cancer. This makes more difficult the identification of false positives based on the results from the CHARGE consortium. The analysis showed that ComBat had the highest level of relative sensitivity, i.e. relatively less false negative CpG associated to smoking, compared to the residuals and SVA, consistently for β or M values. On the other hand, SVA came across as the method with, by far, the highest specificity, possibly indicating lesser predisposition to the commit of false positives. As SVA made a much more aggressive correction of systematic variability, the sites identified by SVA are more likely to be universal disruption due to smoking which can explain its higher specificity and its lower sensitivity. In order to avoid over-adjustment using SVA, latent covariates related to subgroups such as the chip position should not be included in the regression model. SVA outperformed both the residuals and, in particular, ComBat, whose lack of specificity turned out to be substantial. In research domains characterized by the danger of populating the scientific literature with false positive findings, like in the *-omics* era, the performance of SVA towards conservative results was deemed to be a valuable feature. Our results would need to be replicated in another dataset.

The β values are approximations of the percentage of methylation in a CpG site. Their distribution is often skewed and ranged from 0 to 1. On the other hand, M values approximate a normal distribution but are more complex to interpret, as they do not have an obvious biological meaning. It has been recommended to use M values for conducting methylation analysis and to use the β values when reporting results due to their intuitive biological interpretation [26]. In our study, the PC-PR2 method identified the same sources of variability explaining a similar amount of the total variability using M or β values. This is likely a consequence on the fact that PC-PR2 is a descriptive method that does not use statistical inference. The association between smoking and DNA methylation was slightly attenuated in terms of number of significant sites using the M values, rather than β values, for the standard adjustment, residuals' correction and ComBat correction. Only SVA identified more significant sites with the M values. β values were more sensitive but less specific than M values, i.e. more significant sites, including both true and false positive sites.

Approaches for correcting batch effects have been compared using microarray data of gene-expression profiles [33]. In that study, a parametric prior ComBat and a non-parametric ComBat were compared to SVA and to three other methods, including distance-weighted discrimination [34], mean-centering [35] and geometric ratio-based [36] methods. Using two microarray datasets

from brain RNA samples and two simulated datasets, ComBat outperformed overall the other methods. In particular, both parametric and non-parametric ComBat algorithms allowed a better control of the variation attributed to batch effect and a better increase of Pearson's correlation coefficient of the replicates in the microarray data and determined the largest AUC in their assessment of overall performance.

ComBat has also been compared to six other methods to correct for batch effect in microarray data [37], including Deming regression [38], Passing-Bablok regression [39], linear mixed model, a third-grade polynomial regression, the non-linear Qspline method [40] and the ReplicateRUV approach [41]. The first five methods calculate residuals based on different regression models. ReplicateRUV removes unwanted variation based on negative control genes and sample replicates. The combination of quantile normalization and ComBat in large-scale gene expression data in the Gutenberg Health Study removed batch effect and preserved biological variability [37].

In this work, we chose to focus on the residuals, ComBat and SVA approaches, because they are the currently most common methods used to remove unwanted variation in DNA methylation. This work can also be applied to the newer methods which are recently available such as the Bacon approach, a Bayesian method to control bias and inflation in EWAS and TWAS based on estimation of the empirical null distribution [42].

Conclusions

Our results suggest that in order to reduce the contribution to systematic variation of DNA methylation, it is essential to randomly allocate samples within chips and batches. This is particularly relevant in nested studies for case-control pairs, possibly within the same row position within a chip. We have shown that the PC-PR2 method on DNA methylation levels lent itself as a very useful tool to explore an a priori list of laboratory factors and sample characteristics and to identify the ones possibly determining unwanted variability in large-scale dimension sets such as epigenetics data. This step turned out to be essential to guide the choice of correcting methods, such as the regression-based residuals, ComBat or SVA, and to further appreciate the extent of these corrections. These steps should be part of the pre-processing analysis of any *-omics* data. SVA should specifically be considered when sources of variability are not known. ComBat and the residuals' method require that potential sources of variability are identified.

Additional files

Additional file 1: Figure S1. Box plots of global methylation (*M* values) according to laboratory factors: batch (a), chip position within batches (b), sample position within chips (c). (PDF 99 kb)

Additional file 2: Figure S2. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the β values. The inflation factor λ is defined as the ratio of the median of the observed $\log_{10}$ transformed p values from the CpG site-specific analysis and the median of the expected $\log_{10}$ transformed p values. (PDF 110 kb)

Additional file 3: Figure S3. Quantile-quantile (QQ) plots for CpG site-specific analysis with respect to smoking using standard adjustment (a), residuals (b), ComBat (c) and SVA (d) correcting methods for the *M* values. The inflation factor λ is defined as the ratio of the median of the observed $\log_{10}$ transformed p values from the CpG site-specific analysis and the median of the expected $\log_{10}$ transformed p values. (PDF 110 kb)

Abbreviations

BC: Breast cancer; EPIC: European Prospective Investigation into Cancer and nutrition; FDR: False discovery rate; HM450K: Illumina Infinium HumanMethylation450K; PC-PR2: Principal component partial R-square; SVA: Surrogate variables analysis

Acknowledgements

The authors would like to thank the financial support provided by La Fondation de France for a doctoral fellowship. They are also grateful for all the women who participated in the EPIC cohort and without whom this work would not have been possible.

Funding

This work was supported by 'Fondation de France' (2015 00060737) through a doctoral fellow to FP. A grant from the Institut National du Cancer (INCa, France) (2012-070) was awarded to IR and ZH. ZH was also supported by the European Commission (EC) Seventh Framework Programme (FP7) Translational Cancer Research (TRANSCAN) Framework, the Fondation Association pour la Recherche contre le Cancer (ARC, France) and the EC FP7 EurocanPlatform: A European Platform for Translational Cancer Research (grant number: 260791). In addition, this study was supported by postdoctoral fellowship to SA from the International Agency for Research on Cancer, partially supported by the EC FP7 Marie Curie Actions – People – Co-funding of regional, national and international programmes (COFUND). Swedish Cancer Society, Swedish Research Council and County Councils of Skåne and Västerbotten supports SH. AC and KKO are supported by MRC programme grants [MC_UU_12015/1, MC_UU_12015/2 and [MR/L00002/1]. THN is supported by UiT - the Arctic University of Norway. The Hellenic Health Foundation is supporting EPIC-Greece. The funders of the study had no role in study design, data collection, data analysis, data interpretation or writing of the manuscript.

Availability of data and materials

Not applicable.

Authors' contributions

FP performed the statistical data analysis and drafted the manuscript. PF developed the concept of the study with FP, and contributed to draft the manuscript. SA was responsible for the technical aspects of DNA methylation acquisition. IR and ZH conceived the epigenetics study in the nested case-control study on breast cancer, and critically reviewed the manuscript. SA, AK and AN contributed to the interpretation of the results. LB and PV were involved in the data interpretation. All authors contributed to draft the final versions of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study was approved by the Ethical Review Board of the International Agency for Research on Cancer, and by the local Ethics Committees in the participating centres. This study was also conducted in accordance with the IARC Ethic Committee (Project No 10-22).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Nutritional Methodology and Biostatistics Group, International Agency for Research on Cancer (IARC), World Health Organization, 150 cours Albert Thomas, 69372 Lyon CEDEX 08, France. [2]Epigenetics Group, IARC, Lyon, France. [3]Department of Clinical and Experimental Medicine, University of Pisa, Pisa, Italy. [4]MAP5 – UMR CNRS 8145, Université Paris Descartes, Sorbonne Paris Cité, Paris, France. [5]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [6]Department of Radiation Sciences, Oncology, Umeå University, Umeå, Sweden. [7]MRC Epidemiology Unit, Institute of Metabolic Science, University of Cambridge School of Clinical Medicine, Cambridge, UK. [8]IIGM – Italian Institute for Genomic Medicine, Torino, Italy. [9]Department of Community Medicine, UiT - The Arctic University of Norway, Tromsø, Norway. [10]Section for Epidemiology, Department of Public Health, Aarhus University, Aarhus, Denmark. [11]Department of Cardiology, Aalborg University Hospital, Aalborg, Denmark. [12]CESP, Fac. de médecine - Univ. Paris-Sud, Fac. de médecine – UVSQ, INSERM, Université Paris-Saclay, Villejuif, France. [13]Gustave Roussy, Villejuif, France. [14]Centre for Health Protection (pb12), National Institute of Public Health and the Environment (RIVM), Bilthoven, Netherlands. [15]Hellenic Health Foundation, Athens, Greece. [16]WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Athens, Greece. [17]Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain. [18]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [19]MRC/PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [20]Nutritional Epidemiology Group, IARC, Lyon, France.

Received: 22 September 2017 Accepted: 12 March 2018
Published online: 21 March 2018

References

1. Berger SL, Kouzarides T, Shiekhattar R, Shilatifard A. An operational definition of epigenetics. Genes Dev. 2009;23:781–3.
2. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 2016;8:599–618.
3. Guida F, Sandanger TM, Castagne R, Campanella G, Polidoro S, Palli D, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. Hum Mol Genet. 2015;24:2349–59.
4. Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, Mandaviya PR, et al. Epigenetic signatures of cigarette smoking. Circ Cardiovasc Genet. 2016;9:436–47.
5. Kruman II, Fowler AK. Impaired one carbon metabolism and DNA methylation in alcohol toxicity. J Neurochem. 2014;129:770–80.
6. Joubert BR, den Dekker HT, Felix JF, Bohlin J, Ligthart S, Beckett E, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. Nat Commun. 2016;7:10577.
7. Ba Y, Yu H, Liu F, Geng X, Zhu C, Zhu Q, et al. Relationship of folate, vitamin B12 and methylation of insulin-like growth factor-II in maternal and cord blood. Eur J Clin Nutr. 2011;65:480–5.
8. Barrow TM, Michels KB. Epigenetic epidemiology of cancer. Biochem Biophys Res Commun. 2014;455(1-2):70–83. https://doi.org/10.1016/j.bbrc.2014.08.002
9. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–95. https://doi.org/10.1016/j.ygeno.2011.07.007
10. Heyn H, Li N, Ferreira HJ, Moran S, Pisano DG, Gomez A, et al. Distinct DNA methylomes of newborns and centenarians. Proc Natl Acad Sci U S A. 2012;109(26);10522–7. https://doi.org/10.1073/pnas.1120658109
11. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 2011;13:97–109.
12. Herceg Z, Ghantous A, Wild CP, Sklias A, Casati L, Duthie SJ, et al. Roadmap for investigating epigenome deregulation and environmental origins of cancer. Int J Cancer. 2018;142(5):874–82. https://doi.org/10.1002/ijc.31014
13. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11 https://doi.org/10.1038/nrg2825.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics (Oxford, England). 2007;8:118–27.
15. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3. https://doi.org/10.1371/journal.pgen.0030161
16. Leek JT, Storey JD. A general framework for multiple testing dependence. Proc Natl Acad Sci U S A. 2008;105:18718–23.
17. Fages A, Ferrari P, Monni S, Dossus L, Floegel A, Mode N, et al. Investigating sources of variability in metabolomic data in the EPIC study: the principal component partial R-square (PC-PR2) method. Metabolomics. 2014;10:1074–83.
18. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. 2002;5:1113–24.
19. Wang SC, Petronis A. DNA methylation microarrays: experimental design and statistical analysis. Boca Raton: Hall; 2008.
20. Matejcic M, de Batlle J, Ricci C, Biessy C, Perrier F, Huybrechts I, et al. Biomarkers of folate and vitamin B12 and breast cancer risk: report from the EPIC cohort. Int J Cancer. 2017;140:1246–59.
21. Ambatipudi S, Horvath S, Perrier F, Cuenin C, Hernandez-Vargas H, Le Calvez-Kelm F, et al. DNA methylome analysis identifies accelerated epigenetic ageing associated with postmenopausal breast cancer susceptibility. Eur J Cancer. 2017;75:299–307.
22. Illumina. GenomeStudio/BeadStudio software methylation module. 2011.
23. Chen YA, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013;8:203–9.
24. Bolstad BM. Probe level quantile normalization of high density oligonucleotide array data. 2001.
25. 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 (Oxford, England). 2013;29:189–96.
26. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010;11:587.
27. 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;13:1–16.
28. Gower JC. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika. 1966;53:325–38.
29. Kleinbaum DG, Kupper LL, Nizam A, Rosenberg ES. Applied regression analysis and other multivariable methods. Nelson Education; 2013.
30. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, et al. Orchestrating high-throughput genomic analysis with bioconductor. Nat Methods. 2015;12:115–21.
31. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics (Oxford, England). 2012;28:882–3.
32. Wilhelm-Benartzi CS, Koestler DC, Karagas MR, Flanagan JM, Christensen BC, Kelsey KT, et al. Review of processing and analysis methods for DNA methylation array data. Br J Cancer. 2013;109:1394–402.
33. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011;6:e17238.
34. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, et al. Adjustment of systematic microarray data biases. Bioinformatics (Oxford, England). 2004;20:105–14.
35. Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, et al. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. BMC Med Genet. 2008;1:1–14.
36. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. 2010;10:278–91.
37. Müller C, Schillert A, Röthemeier C, Trégouët D-A, Proust C, Binder H, et al. Removing batch effects from longitudinal gene expression—quantile

normalization plus ComBat as best approach for microarray transcriptome data. PLoS One. 2016;11:e0156594.
38. Martin RF. General deming regression for estimating systematic bias and its confidence interval in method-comparison studies. Clin Chem. 2000;46:100–4.
39. Passing H, Bablok W. A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in clinical chemistry, part I. Journal of clinical chemistry and clinical biochemistry. Zeitschrift fur klinische Chemie und klinische Biochemie. 1983;21:709–20.
40. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielser HB, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 2002;3:research0048.
41. Jacob L, Gagnon-Bartsch JA, Speed TP. Correcting gene expression data when neither the unwanted variation nor the factor of interest are observed. Biostatistics (Oxford, England). 2016;17:16–28.
42. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol. 2017;18:19.

Briefings in Bioinformatics, 00(0), 2019, 1–12

doi: 10.1093/bib/bbz015
Advance Access Publication Date: 14 February 2019
Review article

Evaluation of integrative clustering methods for the analysis of multi-omics data

Cécile Chauvel*, Alexei Novoloaca*, Pierre Veyre, Frédéric Reynier and Jérémie Becker

Corresponding author: Jérémie Becker, BIOASTER Research Institute, 40 avenue Tony Garnier, 69007 Lyon, France. Tel.: +33 4 69 85 19 21; Fax: +33 4 72 70 48 2; E-mail: jeremie.becker@bioaster.org
*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Abstract

Recent advances in sequencing, mass spectrometry and cytometry technologies have enabled researchers to collect large-scale omics data from the same set of biological samples. The joint analysis of multiple omics offers the opportunity to uncover coordinated cellular processes acting across different omic layers. In this work, we present a thorough comparison of a selection of recent integrative clustering approaches, including Bayesian (BCC and MDI) and matrix factorization approaches (iCluster, moCluster, JIVE and iNMF). Based on simulations, the methods were evaluated on their sensitivity and their ability to recover both the correct number of clusters and the simulated clustering at the common and data-specific levels. Standard non-integrative approaches were also included to quantify the added value of integrative methods. For most matrix factorization methods and one Bayesian approach (BCC), the shared and specific structures were successfully recovered with high and moderate accuracy, respectively. An opposite behavior was observed on non-integrative approaches, i.e. high performances on specific structures only. Finally, we applied the methods on the Cancer Genome Atlas breast cancer data set to check whether results based on experimental data were consistent with those obtained in the simulations.

Key words: benchmark; clustering; data integration; multi-omics; unsupervised analysis

Introduction

The accumulation of large molecular data sets has fueled the development of translational bioinformatics and systems biology that share a holistic view on omics data. While the former aims to link biological to clinical data to improve our understanding of disease mechanisms, the latter explores the basic functional properties of living organisms based on the premise that biological processes build upon the interplay between macromolecules. Both approaches rely on the idea that biological mechanisms (and, more generally, phenotypic traits) can only be fully captured through the study of molecular interactions among different omics layers.

Multi-omic approaches have received much attention in recent years for their potential applications in clinics. In genome-wide association studies for example, the mechanisms by which the identified loci influence phenotypes remain generally unknown and are likely to be unveiled using functional

Cécile Chauvel, PhD, is a researcher in biostatistics in the Data Management and Analysis unit at Bioaster, Lyon, France.
Alexei Novoloaca is a PhD student in biostatistics in the Epigenetics Group at the International Agency for Research on Cancer, World Health Organization, Lyon, France.
Pierre Veyre is a computer scientist in the Data Management and Analysis unit at Bioaster, Lyon, France.
Frédéric Reynier, PhD, is the head of the Genomics and Transcriptomics unit at Bioaster, Lyon, France.
Jérémie Becker, DPhil, is a researcher in biostatistics in the Genomics and Transcriptomics unit at Bioaster, Lyon, France. BIOASTER is a technological research institute in microbiology that aims to develop new innovative and high-value technology solutions through collaborative projects. Its main interest lies in tackling antimicrobial resistance, developing new diagnostics, improving vaccines' safety and efficacy and understanding the involvement of microbiome in human and animal health.
Submitted: 26 September 2018; **Received (in revised form):** 12 January 2019

genomics. Cancer subtype diagnosis, commonly determined by clinicopathologic parameters (i.e. morphological variables), tend to underestimate inter-patient variability by classifying patients with different responses to treatment or long-term prognosis in the same group [1]. By crossing genomic, epigenomic, transcriptomic and proteomic data, the Cancer Genome Atlas (TCGA) network was able to refine breast cancer classes into phenotypically homogeneous groups [2]. Quigley *et al.* [3] demonstrated that genetic susceptibility in breast cancer is, in some cases, context specific and requires the combination of transcriptomic and epigenomic data to explain the mechanisms of risk alleles. Similarly, Meng *et al.* [4] showed that the leukemia extravasation signaling pathway could only be identified through the integration of gene and protein expression. These observations show that the integration of multiple sources has the potential to (i) mitigate the risk of false positives when multiple sources of evidence point to the same pathway [5], (ii) lead to novel insights into the molecular crosstalk between omics layers that underlies complex traits and (iii) identify new biomarkers to stratify patients into novel, clinically relevant disease subtypes [6].

The increasing availability of large heterogeneous data sets (e.g. TCGA, the International Cancer Genome Consortium and the Asian Cancer Research Group) has prompted the development of novel integrative methods that aim to capture weak yet consistent patterns across data types. This task is, however, non-trivial due to (i) the increased dimensionality that makes inference weaker, (ii) the challenge to decipher data-specific from inter-source variations and (iii) the different types of noise and confounding effects across platforms, resulting in data heterogeneity. For example, next-generation sequencing (NGS) and microarray data are commonly modeled with negative binomial and Gaussian distributions, respectively.

Despite these challenges, at least five major strategies have been proposed to integrate heterogeneous omics data. The first strategy, conceptual integration, consists of analyzing each omics separately and combining the results at the interpretation step. Because of its simplicity and the lack of gold standard in the domain, this type of integration has been largely applied in multi-omic analysis [7]. One obvious drawback of such method is its limited power to uncover modest but coordinated variations acting at different biological layers [8]. The 2nd strategy, consensus clustering, generates an overall sample classification after an initial clustering step performed in each omics. Although successfully applied in TCGA to refine breast cancer subtypes [2], this two-step procedure of separate clusterings followed by *post hoc* integration limits the power to detect crosstalk between omics. The 3rd strategy, concatenation-based integration, allows the application of standard machine learning techniques after concatenation of omics measurements into a single matrix. While such strategy turned out to have high discriminative power in supervised framework [9], it is sensitive to the data size when applied naively and consequently returns results biased toward the omics with the most numerous features. However, recent supervised concatenation approaches account for unbalanced data sizes [10]. Additionally, concatenation-based integration does not account for relationships across sources and heterogeneous measurement error across platforms. The 4th strategy searches for common variations across omics using matrix factorization, Bayesian and network-based approaches specifically tailored for data integration. The 5th strategy, multi-omic pathway enrichment, aims to find pathways that correlate with a particular phenotypic end point, based on their multi-omic profiles. In practice though, the current tools perform pathway enrichment in each omics before combining the obtained *P*-values [11, 12], similarly to the conceptual integration.

Most of the effort in the area have been concentrated on the 4th strategy where many methods propose to find a joint cluster structure, from which patient stratification and molecular mechanisms can be deduced. The methodological aspects underlying integrative approaches have been recently reviewed [13–16] and led to a classification according to two criteria; whether or not the method under consideration relies on (i) networks and (ii) Bayesian approaches, the network-free non-Bayesian approaches being based on matrix factorization [17].

Most of these methods have been evaluated individually and occasionally compared against iCluster (presented hereafter). To our knowledge, only one benchmark of five network-based and matrix factorization approaches has been performed so far [18], leaving Bayesian methods aside. The methods were evaluated with their default parameters, except for the best-performing one.

In the present work, we propose a comparison of six popular methods, one being in common with [18], using simulated and real-world data (TCGA). Because we are interested in clustering approaches that do not require any prior biological knowledge and that can be widely applicable, the methods were selected on the basis of their ability to deal with any data type and produce clustering at both molecular and sample levels. Therefore, models tailored for specific omic types [19–21] and network-based approaches focusing on patient stratification [22] or network enrichment [23] were left aside.

The methods assessed in this work fall into two categories, Bayesian approaches that extend the finite Dirichlet mixture model and dimension reduction techniques aiming at identifying shared latent variables. To avoid favoring one category over the other in our study, the simulations were generated using one model from each family. Also, like any clustering problem, the determination of the optimal number of clusters is crucial and needs to be addressed carefully. To do so, all methods propose guidelines (presented in their description hereafter) and, in most cases, include the associated code. Because (i) this step has a large impact on the final clustering, (ii) the code is not systematically available and, (iii) in practice, users often test different number of clusters and validate their choice using orthogonal data [35], we decided to separately evaluate this step of estimation of the number of clusters from the clustering itself. In this latter evaluation of the clustering step, the methods were run using the true (simulated) number of clusters and evaluated on their ability to recover the simulated clustering.

The remainder of the paper is organized as follows: in the first section, we briefly present the methods, the simulation scenarios and the evaluation criteria. In the Results section, we present the relative performances of the methods both on simulated and TCGA data. Finally, in the light of the results, we discuss the choice of methods in multi-omic framework in Discussion and conclusion.

Methods

Methods overview

iCluster is a Gaussian joint latent variable model that seeks a single-shared clustering structure across K data sets X_k of dimensions $p_k \times N$ ($k = 1, \cdots, K$) measured on the same N samples [24]. Its formulation relies on a latent variable model that captures correlations among variables through latent factors.

iCluster jointly fits K such models with the constraint that the latent variable matrix is shared across data sets:

$$X_k = W_k Z + \varepsilon_k, \qquad Z \sim \mathcal{N}_q(0, I), \tag{1}$$

where W_k is the $p_k \times q$ loading matrix associated with data set k, Z is the $q \times N$ common latent variable matrix and ε_k is the $p_k \times N$ uncorrelated error matrix that follows a multivariate Gaussian distribution $\mathcal{N}_{p_k}(0, \psi_k)$ with zero mean and diagonal covariance matrix $\psi_k = \text{diag}(\sigma^2_{k,1}, \cdots, \sigma^2_{k,p_k})$. By forcing the latent variables to be equal, iCluster assumes that the K data sets derive from a set of common factors. Parameter estimation is then performed using an expectation–maximization algorithm on the multivariate normal distribution. The final cluster assignment is determined by k-means clustering on the posterior expectation of the latent factors $E(Z|X)$. An l_1 penalty can be imposed on the loading coefficients to perform variable selection. The penalty parameter and the number q of latent variables are tuned manually using the proportion of deviance (a measure of cluster separability, [25]) from which the number of clusters can be deduced. Indeed, Shen *et al.* [25] recall that a $K - 1$ rank matrix is sufficient for separating K clusters. In Shen *et al.* [24], a cross-validated Rand index (RI) (Clustering performance criteria), measuring the clustering similarity between the training and the test sets, is used for parameter tuning [24]. An extension of the model, iClusterPlus (not evaluated in the present work), allows to account for binary, counts and categorical data.

moCluster also decomposes each data set X_k into a product of shared latent variables Z and a sparse, data-specific, loading matrix W_k [26], similarly to iCluster. The main difference between these two methods is that iCluster derives from factor analysis, whereas moCluster relies on consensus PCA. For this reason, iCluster separates the covariance from the variable-specific variance, allowing heteroscedasticity among omic features. Consensus PCA, on the other hand, assumes that the noise has same variance across variables ($\sigma^2_{k,j} := \sigma^2$, for all $j = 1, \ldots, p_k$ and $k = 1, \ldots, K$), making common and unique variations no longer separable [24]. Although this assumption is strong for the analysis of heterogeneous omics, moCluster offers a 100–1000-fold speed increase as compared to iCluster due to its deterministic parameter estimation based on the NIPALS algorithm. A step of soft thresholding may be used for variable selection. To determine the number of latent variables, the authors suggest a visual inspection of the eigenvalues (scree plot) or a permutation test on the same eigenvalues. Similarly, the authors propose to perform a hierarchical clustering on the latent variable matrix to obtain the final clustering, the number of clusters being determined with the gap statistic [27].

JIVE extends iCluster and moCluster by adding a data-specific term [28]. This improvement is motivated by the biological interest of studying individual structures and also by observing that data-specific variations can dramatically impact the estimation of the shared structure in partial least squares models [29]. Again, each term factorizes into a loading and a latent variable matrix:

$$X_k = W_k Z + W^s_k Z^s_k + \varepsilon_k, \tag{2}$$

where W^s_k of size $p_k \times q_k$ and Z^s_k of size $q_k \times N$ are the data-specific loading and latent variable matrices, respectively. Note that q and q_k are not necessarily equal, implying that the joint and individual low-rank approximations may be of different dimensions. To guarantee the identifiability of the decomposition, the authors imposed an orthogonality constraint between the joint and individual terms. The parameter estimation is performed by estimating the joint and individual structures iteratively via SVD by fixing one term at a time and minimizing the square norm of the residual matrix for updating the other term. Sparsity is induced during the estimation procedure by an l_1 penalty on the loading matrices. The level of sparsity is determined using the Bayesian information criterion. As with moCluster, the number of joint and individual latent variables is estimated using a permutation approach on the eigenvalues. Unlike iCluster and moCluster that provide tools to cluster samples from the latent variables, Lock and Dunson [35] do not give guideline to generate a final sample clustering.

Similarly to JIVE, iNMF aims to capture the shared and data-specific structures with, however, two notable differences [30]. First, the latent variables are estimated using a non-negativity constraint instead of orthogonality. Second, a coefficient matrix W_k is shared between the data-specific Z^s_k and the common Z basis matrices where the coefficient and basis matrices are the counterparts of the loading and latent variable matrices. iNMF optimizes the following problem with a Euclidean loss function:

$$\min_{\substack{Z, Z^s_1, \cdots, Z^s_K \\ W_1, \cdots, W_K}} \sum_{k=1}^{K} ||X_k - (Z + Z^s_k) W_k||^2 + \lambda \sum_{k=1}^{K} ||Z^s_k W_k||^2. \tag{3}$$

Again, the authors motivate the addition of data-specific effects by demonstrating that jNMF [31], a similar approach without data-specific term, is more sensitive to random noise and confounding effects. The choice of non-negativity factorization is, on the other hand, motivated by its simple and meaningful interpretation that 'the whole is an additive linear combination of its parts' [32]. While non-negative factorization approaches have a naturally sparse and parts-based representation [33], sparsity is nevertheless induced in iNMF by applying an l_1-penalization on the data-specific term. This constraint imposed on the data-specific effects implies that the parameter λ controls for the factorization homogeneity. The authors also propose to apply an l_1-penalty on the coefficient matrix W_k to enforce variable selection. The dimension of the shared and specific structures (equal due to the shared coefficient matrix) is chosen through a consensus-based approach that selects the number of latent variables maximizing the clusters stability across multiple iNMF runs. The λ parameter is, on the other hand, determined using an *ad hoc* procedure that aims to attribute as much of the data as possible to the specific structure while controlling for overfitting. Unlike the previous approaches, the authors propose a method to perform clustering on the variables, which is out of the scope of this work. Similarly to JIVE, no guidelines are provided to obtain a final sample clustering.

Multiple data set integration (MDI) is a Bayesian method that represents each data set k with a Dirichlet-multinomial allocation mixture model [34]. Such mixture model has gained increased popularity for the flexibility offered by the dependency structure, and the different parametric forms the mixture components can adopt. The originality of MDI arises from the way it captures the common structures through pairwise dependencies between data type clusterings. The sample assignment in data set k can thus influence the sample assignment in data set l, allowing the identification of samples that tend to

cluster together in one, some or all data sets. This feature is an important improvement over the previous approaches that assume that the shared structure has to be common across all data sets. The association among data sets is expressed at the level of the component allocation variables with the conditional prior

$$P(c_{i1}, c_{i2}, \cdots c_{iK}|\phi) \propto \prod_{k=1}^{K} \pi_{c_{ik}k} \prod_{k=1}^{K-1} \prod_{l=k+1}^{K} (1 + \phi_{kl} \mathbb{1}(c_{ik} = c_{il})), \quad (4)$$

where $\mathbb{1}$ is the indicator function, ϕ_{kl} controls the association strength between data sets k and l, c_{ik} indicates the cluster allocation of sample i in data set k and $\pi_{c_{ik}k}$ is the mixture proportion associated with cluster c_{ik} in data set k. The parameters, including the number of clusters in each source, are inferred via Gibbs sampling. The authors then propose to maximize the posterior expected adjusted Rand index (PEAR) across source-specific clusterings to determine a single global clustering. As pointed out by Wei *et al.* [14], the model could be extended by modeling the pairwise association at the component level instead of the data set level. Variable selection is not provided by the method, and the maximal number of clusters needs to be fixed by the user. For computational reasons, the authors recommend to set this parameter to half the sample size. However, we noticed that this value led to numerical instabilities and set it to the sample size.

Bayesian consensus clustering (BCC) also extends the Dirichlet mixture model [35]. However, instead of modeling cluster dependency through pairwise association between sources, it aims at uncovering a single common clustering across sources, similarly to the matrix factorization approaches. This is achieved by relating the source-specific clustering L_k in data set k to a consensus clustering through the following dependence function:

$$P(L_{kn} = l|C_n) = \begin{cases} \alpha_k \text{ if } C_n = l \\ \frac{1-\alpha_k}{1-q} \text{ otherwise,} \end{cases} \quad (5)$$

where, for sample n, C_n and L_{kn} are the overall and source-specific cluster allocations in data source k, α_k is the adherence of data set k to the overall clustering and q is the maximum number of clusters (both shared and source specific). The adherence parameter α_k models how intertwined specific and shared clusters are. The parameter q is chosen so that the mean adherence over the sources is maximized, which, according to the authors, results in a small number of selected clusters. Similarly to MDI, a Gibbs sampler is used to estimate the posterior distribution of the parameters.

Data pre-processing

Depending on the model assumptions, the six methods propose different pre-processing steps. Since moCluster and JIVE rely on techniques that treat covariance and variance identically (consensus PCA and SVD), these methods are sensitive to variable scaling. For this reason, data sets are centered or standardized in JIVE and moCluster, respectively. To circumvent the case where 'the largest data set wins', data matrices are further weighted by the reverse of their first eigenvalue (moCluster) or their Frobenius norm (JIVE). iNMF also normalizes each matrix by its Frobenius norm after variance stabilization (log transformation) and non-negativity transformation. By contrast, given that iCluster allows heteroscedasticity, only a centering step is performed. At last, no pre-processing is performed with MDI and BCC, considering that Dirichlet mixture models offer enough flexibility. A brief description of the methods, their pre-processing and availability are provided in Table 1.

Simulation scenarios

The methods presented above were evaluated both on simulated and real data. On simulations, the methods were evaluated on their ability to (i) recover the number of simulated clusters and identify the correct clustering at the (ii) common and (iii) data-specific levels as well as on their (iv) sensitivity. The sensitivity was assessed by varying the level of signal-to-noise ratio (SNR) and the dimension of the shared clusters. Overlaps between the two structures were introduced to assess whether they could hinder the identification of the shared structure. Since the tested methods can roughly be divided in two groups (matrix factorization and Bayesian models), simulations were generated under iNMF (matrix factorization) and BCC (Bayesian model) models to ensure an unbiased evaluation. Each simulation consists of $K = 3$ matrices X_k of dimension $p_k \times N$, with $N = 60$ samples, and p_k features in the 3 data matrices with $P = (180, 210, 240)$. Each matrix X_k consists of 3 common clusters made of 20 samples each. Although the number of features is at least one order of magnitude smaller than what is commonly observed in high-throughput omics, they are more amenable to the present large-scale evaluation in terms of runtimes. To evaluate the validity of our results under more realistic settings though, i.e. higher dimension and an important unbalanced number of features, one scenario was also generated with $P = (300, 600, 3000)$ features. Details and illustrations of the following simulation scenarios are provided in Supplementary Materials. In addition, the runtimes are provided in Supplementary Table 2.

iNMF-derived scenarios

The first simulation scenarios are derived from iNMF, in which each data matrix X_k is built as the sum of three matrices: one made of three shared diagonal blocks of same dimensions, one with one or two data-specific off-diagonal blocks and one made of random uniform noise. The blocks were constructed by multiplying the binary latent variables (Z and Z_k) with the data-specific loadings W_k. The loadings were simulated under a beta(2,2) distribution, satisfying thus the non-negativity constraint required by iNMF while not diverging too much from a Gaussian distribution (the beta and Gaussian distribution are symmetrical and have a bell shape). The same two levels of noise used in [30] were also used here. This level of noise is controlled by a 'scattered error' that replaces either a positive value with zero or a zero with a randomly generated $(beta(2, 2) \times 2)^2$ with a probability $1 - \sigma_s$ dependent of the desired level of noise. A distinctive feature of the simulations in [30] is that the data-specific blocks are aligned with the columns (variables) of the shared structure. In practice, this premise implies that features involved in shared and omic-specific mechanisms are identical. Because probably unrealistic, we simulated specific blocks so they do or do not overlap with the shared structure, where in the first case, the specific blocks randomly overlap with one or two shared block(s). In the context of our study, an overlap between a common and a specific block means that they have some variables in common (Supplementary Figure 1). For both overlap

Table 1. Description, pre-processing and implementation of the evaluated methods

	Method	Description	Pre-processing	Implementation
Integrative methods	iCluster	Joint latent variable model	Centering	R package iCluster
	moCluster	Modified consensus PCA	Standardization	R package mogsa
	JIVE	Matrix factorization into common and specific variations	Standardization	R package r.jive
	iNMF	Joint non-negative matrix factorization	Variance stabilization Non-negativity transformation Frobenius normalization	Python script
	MDI	Dirichlet mixture models	None	Matlab script
	BCC	Dirichlet mixture models	None	R package bayesCC
Non-integrative methods	GMM	GMMs	None	R package mclust
	Concatenation	Concatenation and GMMs	None	R package mclust
	Consensus clustering	GMMs and maximization of PEAR	None	R packages mclust and mcclust

and non-overlap simulations, a given observation belongs to one common cluster and zero to two specific clusters.

Overall, 3 scenarios were generated: iNMF overlap, iNMF non-overlap and iNMF high dimensional, the third being identical to the first apart from the number of features equals to $p = (300, 600, 3000)$. For each combination of scenario and SNR, 100 simulations were generated, adding up to 600 simulations.

BCC-derived scenario

In the same way as the iNMF scenario, the simulation scenario proposed in BCC was extended in two ways: three to five specific clusters adhering loosely to three overall consensus clusters were simulated. Instead of simulating each feature with univariate Gaussian distributions, realizations of p_k-dimensional Gaussian distributions were generated using the MixSim R package. Unlike iNMF simulations, each observation is uniquely assigned to one specific and one shared cluster. Furthermore, the level of SNR is set by the hypercube parameter in MixSim that controls the space in which the cluster means are sampled. Again, 2 levels of noise were tested, for each of which, 100 simulations were generated, resulting in 200 simulations.

Sensitivity scenario

Methods' sensitivity was only evaluated on the iNMF scenario by reducing block sizes (both shared and specific). The matrix dimensions were held constant, which implies that samples outside the shared and specific blocks were generated with noise. The number of samples by shared blocks n_b took values in $\{5, 8, 11, 14, 17, 20\}$. For each combination of noise, overlap (same as in iNMF scenario) and block size, 20 simulations were generated, adding up to a total of 480 simulations.

Clustering performance criteria

The consistency between two clusterings or partitions is commonly measured using the RI [36]. Given c and $\hat{c}$ the simulated and estimated clusterings (containing the cluster assignment for each sample), the RI calculation relies on the classification of each sample pair in one of four possible categories. Let a be the number of sample pairs in the same cluster in c and $\hat{c}$, b be the number of pairs in the same cluster in c but not in $\hat{c}$, c be the number of pairs in the same cluster in $\hat{c}$ but not in c and d be the number of pairs in different clusters in c and $\hat{c}$.

The RI is then defined as

$$RI = \frac{a+d}{a+b+c+d}.$$

For 2 partitions in perfect agreement, the RI is 1. However, because the Rand index expectation of two random partitions is not constant, Hubert and Arabie [37] introduced the adjusted Rand index (ARI) as

$$ARI = \frac{RI - E(RI)}{\max(RI) - E(RI)}, \tag{6}$$

where $E(RI)$ is the expected RI in case of independence between the two partitions. In the following, the measure of agreement between simulated and estimated clusterings will be presented using the ARI.

From Methods overview we recall that for a given simulation, global and data-specific clusterings are obtained for all methods except iCluster and moCluster for which only a common clustering is available.

Results

In the results presented hereafter, sparsity parameters were left aside, which implies that no penalization was applied in any analysis.

Simulated data

In this section, we evaluate (i) the methods' ability to recover the correct number of clusters, (ii) the consistency between simulated and estimated clusterings and (iii) the methods' sensitivity, based on simulated data.

Determination of the number of common clusters

Before evaluating the clustering performances, we first sought to assess the methods' ability to estimate the number of common clusters. To ensure an unbiased comparison, built-in methods proposed by matrix factorization approaches to estimate the number of latent variables and clusters as well as the homogeneity parameter λ in iNMF were run according

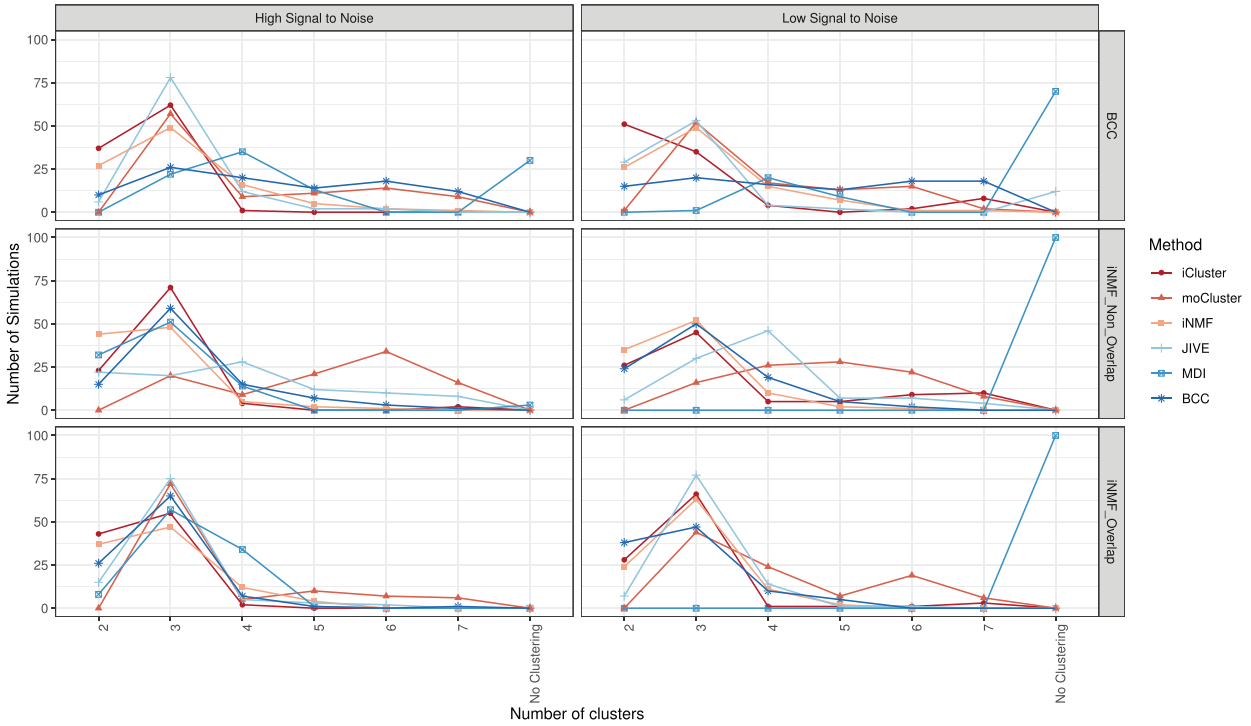


Figure 1. Ability to recover the correct number of simulated clusters: number of shared clusters estimated in each simulation scenario (rows) and SNR (columns).

to the authors' guidelines (for each method, a summary is provided in Methods overview). The other parameters were left at their default values. We recall that the present evaluation is performed at the cluster level, the clusters being generated by all methods except JIVE and iNMF. An additional step of k-means (100 repetitions to ensure stability) on the latent variables was thus added for these 2 methods, the number of clusters being either determined by the gap statistic for JIVE or set to the number of estimated latent variables for iNMF. This choice was motivated by the correspondence made by the authors of iNMF between the number of 'modules' (latent variables) and the number of bi-clusters.

The number of clusters estimated by each method on the 600 simulations is shown in Figure 1. We recall that each simulation consists of three data sets sharing three global clusters, as described in Simulation scenarios. One can first notice that the methods correctly retrieved three clusters on average. The distributions appear sharper around the modes for the high SNR and the iNMF-overlap scenario. Conversely, the methods globally perform poorly on the iNMF-non-overlap scenario, probably because these simulations contain the same number of shared and specific variables, and that common blocks have half as many variables as those in the iNMF-overlap scenario. Continuing with the iNMF-non-overlap scenario, the estimated number of clusters is uniform for moCluster and JIVE, while it successfully peaks around three clusters for iCluster and BCC. These observations suggest that the latter are more robust to the reduction of shared blocks. On BCC simulations, unlike matrix factorization approaches, the two Bayesian methods either do not detect clusters (MDI) or fail to identify the three global clusters (BCC). These results are surprising since one would expect a method run on simulations generated with its own model to perform well. The systematic absence of clustering returned by MDI for low SNR simulations, regardless of the scenario, implies that the method lacks robustness against noise. One can finally note systematic biases in four methods: iNMF and iCluster on the one hand, moCluster and MDI on the other hand suffer from under and over-estimation, respectively. This is particularly true on BCC and iNMF-non-overlap simulations. Overall, iNMF, iCluster, JIVE, BCC, moCluster and MDI successfully recovered three clusters in 62.7, 55.7, 55.5, 44.5, 43.5 and 21.8% of the simulations respectively.

Method performances on shared and specific structures

After evaluating the methods' ability to estimate the number of clusters, we now assess the clustering quality by measuring the coherence between simulated and estimated clusterings both at the shared and data-specific levels. In this analysis, methods were configured so that the expected number of clusters are set to the true number of simulated clusters, except for Bayesian approaches on specific clusters. Indeed, the specific clustering depends on the simulation, matrix and method considered. The dependence on the method exists because specific structures are modeled differently in matrix factorization and Bayesian methods. This modeling difference only arises in iNMF simulations when a specific block overlap with two common blocks; in this situation, matrix factorization methods (JIVE and iNMF) are designed to recover all three blocks (two common and one specific clusters), whereas Bayesian approaches see three specific clusters. Therefore, when computing ARI for specific structures, the expected clustering was provided by unique blocks in Z_k^s or $Z + Z_k^s$ when run with matrix factorization or Bayesian approaches, respectively. Similarly, the number of specific clusters was set to the number of these unique blocks.

Table 2. Best-performing parameters selected by grid search before method comparison. The ranges tested for each parameter are indicated in parenthesis (see Supplementary Figures 2–4)

		JIVE	iNMF	BCC
Number of latent variables / modules		2 (2–5)	3 (2–6)	–
Homogeneity level		–	0.3 (0.01,0.03,0.1,0.3)	–
Number of clusters	iNMF simulations	–	–	3 (2–7 and 'as simulated')
	BCC simulations	–	–	as simulated (2–7 and 'as simulated')

To guarantee a fair comparison across methods, the parameters were set to their best-performing values defined as either the value used in data simulation or the ones maximizing the ARI with the global clustering. Indeed, some parameters were straightforward to set because fixed in the simulations, namely, the number of common and specific clusters as well as the number of latent variables in iCluster and moCluster [rank$(Z) = 2$]. For the others, i.e. the number of latent variables in JIVE, the number of modules and the homogeneity level in iNMF, the maximum number of clusters q in BCC, a grid search aiming at finding the parameters maximizing the ARI was conducted. Certain parameters in iNMF (homogeneity) and BCC (maximum number of clusters) could favor the common or specific structure over the other. Since the goal in data integration is to identify common variations, the parameters were tuned to maximize the ARI of the global clustering.

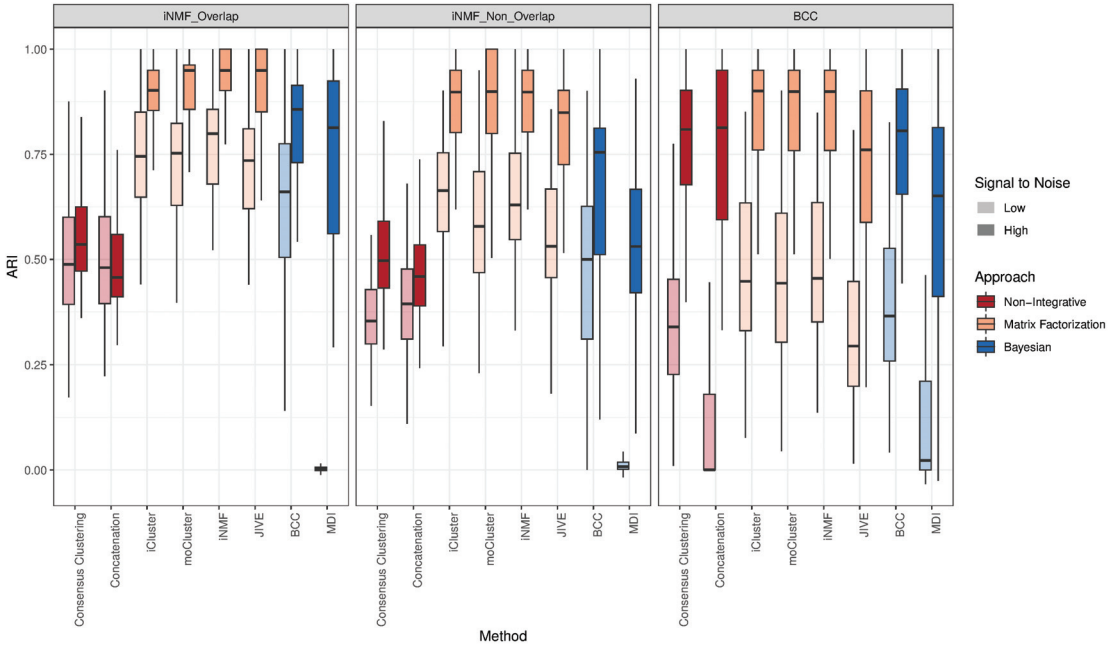
Supplementary Figures 2–4 display the performances of these three methods on the common and specific structures for all simulation scenarios. Starting with JIVE, the rank does not show much effect on the method performances. For the specific structures, ARIs are under 0.25, indicating a poor ability to recover them. For the shared structures, ARIs appear on average slightly higher when rank equals 2, value retained in the method comparison. For iNMF, a clear increase in ARI between two and three modules followed by a plateau led to the selection of three modules. Unsurprisingly, the homogeneity parameter has a large impact on the recovery of the shared or specific structures, trend particularly apparent on BCC simulations where the largest (resp. smallest) homogeneity value allows an almost perfect identification of the common (resp. specific) structures (ARI $\approx$1). The selected homogeneity value was the one favoring most the common structure, i.e. $\lambda = 0.3$. Similarly, Supplementary Figure 4 reveals an important effect of the number of clusters on BCC performances for the first two simulation scenarios: a bell-shaped curve peaking at three clusters is obtained with iNMF simulations, while the highest ARIs are attained when the number of clusters was set to this used in simulations ('As Simulated') for BCC simulations and this for both shared and specific structures. By contrast, the ARI shows almost no variations across cluster numbers on the sensitivity scenario. Those parameter values selected by grid search or from the simulation design are summarized in Table 2.

We now turn to the method comparison at the shared level. In addition to the six integrative methods, two alternatives mentioned in the introduction were included to evaluate the added value of integration: consensus clustering based on maximization of the PEAR [38] and Gaussian mixture model (GMM) clustering on concatenated matrices using the Mclust R package. It can first be noticed that most integrative methods display high ARI, suggesting a good ability to recover common clusters (Figure 2a). The SNR has a large impact on the performances, especially on MDI and the concatenation approach (BCC simulations only), which both lack robustness against noise. Similar trends are observed in overlap and non-overlap-iNMF simulations where matrix factorization approaches have equivalent ARI and outperform Bayesian methods, while non-integrative approaches show smaller ARI. Similarly to the previous section, the performances decrease in iNMF-non-overlap simulations, which can again be attributed to the fact that common blocks contain half as many variables as in the overlap scenario. This drop is more accentuated with MDI, supporting the idea that the method is more sensitive to data perturbations. Looking at BCC simulations, iCluster, moCluster and iNMF outperform again the others, shortly followed by consensus clustering, JIVE and BCC. Matrix concatenation and MDI, on the other hand, show relatively large ARIs when SNR is high but perform poorly at low SNR.

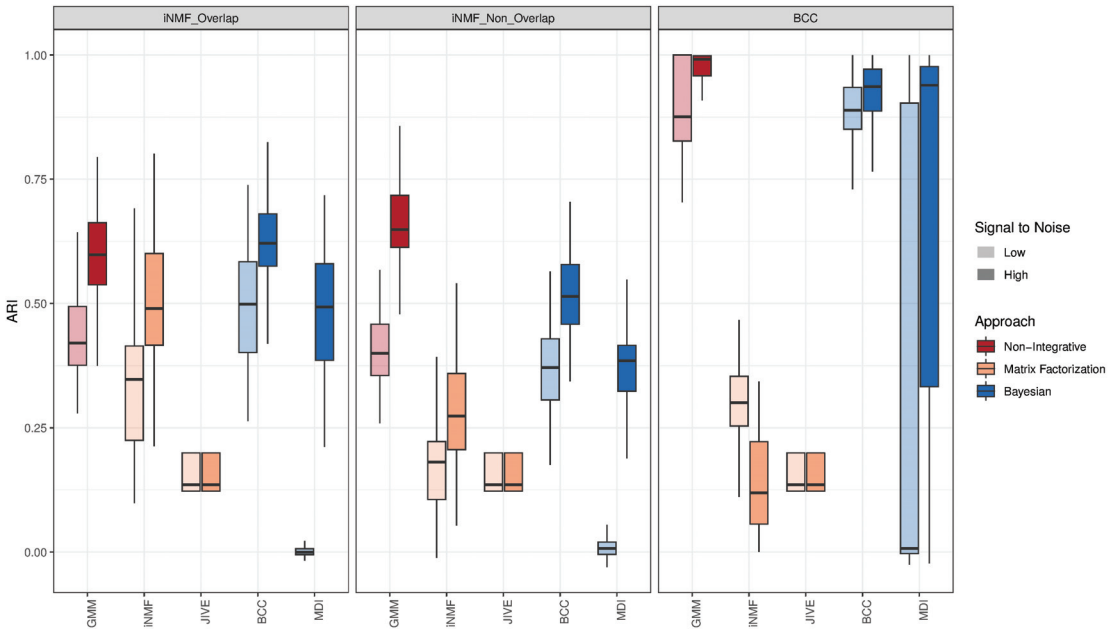
The results with matrix concatenation are in line with the previous study [9]. Overall, matrix factorization methods and, more particularly, iCluster, moCluster and iNMF present the best ability to recover common structures, and, this, regardless of the simulation scenario. BCC shows an intermediate behavior between matrix factorization approaches and MDI. Despite a moderate robustness when shared and specific blocks do not overlap, BCC displays fairly high ARIs on BCC simulations, probably because the simulations were generated from the same model.

The detection of data-specific structures is not central in this study given that standard clustering approaches are tailored for this task. We nevertheless evaluated this functionality because one is generally interested in both shared and specific structures when performing multi-omic studies. Similarly to the evaluation of the shared structures, GMM was added in the comparison to benchmark the four methods against a standard (non-integrative) clustering approach. In the two simulation scenarios, GMM, closely followed by BCC, outperforms the 3 other integrative methods with ARIs close to one in BCC simulations (Figure 2b). Matrix factorization methods, JIVE in particular, performs poorly in all scenarios. In the same way as the shared structure evaluation, MDI achieves close to zero ARI at low SNR confirming its lack of robustness to noise but nevertheless showed intermediate ARI values at high SNR. These results are not unexpected since GMM and BCC are designed focus on data-specific clustering, whereas iNMF and JIVE aim to recover shared clusters.

To conclude on this section, the six methods showed a real improvement over non-integrative approaches to find shared clusters on iNMF simulations, while only iCluster, moCluster and iNMF did so on BCC simulations. By contrast, the methods failed to reach GMM performances on specific clusters, except for BCC. Unsurprisingly, no method could properly identify shared and specific structures simultaneously. Because the detection of either structure is largely influenced by parameters, the latter must be carefully tuned according to the study goals. The same trends were also observed in the iNMF-high-dimensional scenario, which indicates that the results also apply in high



(a) Performances on shared structures



(b) Performances on specific structures

Figure 2. Consistency between simulated and estimated clusterings: ARI boxplots are displayed on shared (a) and specific (b) structures for each simulation scenario (columns) and SNR (transparency).

dimension, when the number of features differs across data sets (Supplementary Figures 5–7).

Evaluation of methods sensitivity

The present study of sensitivity aims at determining whether the methods accurately identify common structures when their size is reduced up to $n_b = 5$ samples per block. No additional tuning step was required as parameters were determined for all scenarios, including the sensitivity ones, in the previous section. Only the number of expected clusters was changed to 4 when $n_b \leq 17$, i.e. when a noise cluster was present.

As already noticed in the previous results, SNR and overlaps across structures largely influence the performances,

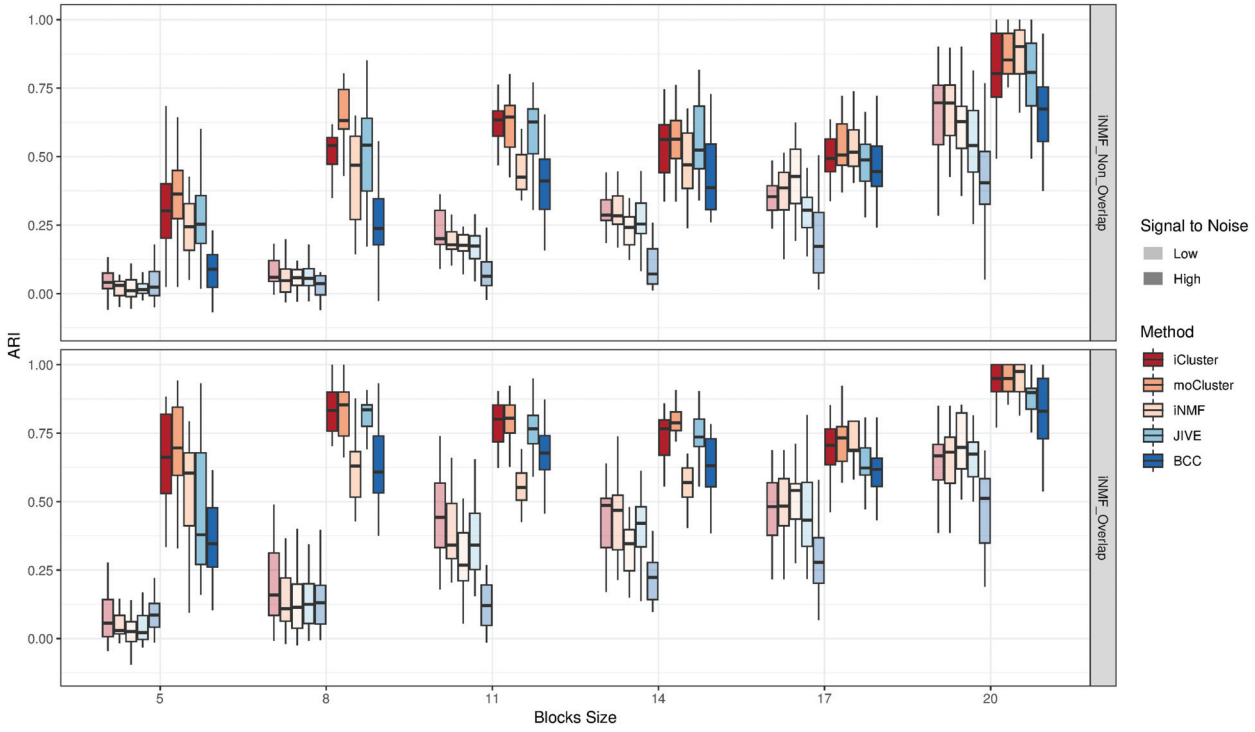


Figure 3. Evaluation of methods' sensitivity: ARI distributions are displayed for different block sizes ($n_b \in \{5, 8, 11, 14, 17, 20\}$), simulation scenarios (rows) and SNRs (transparency).

especially at $n_b = 5$ where ARIs are roughly twice as large in the iNMF-overlap as in the iNMF-non-overlap scenario (Figure 3). As expected, the size of the blocks also impacts the performances, with ARIs increasing with the number of samples per block. Surprisingly, this trend is not linear at high SNR where performances successively increase ($n_b \in \{5, 8\}$), plateau or slightly drop ($n_b \in \{8, 17\}$) and increase again ($n_b \in \{17, 20\}$). To rule out possible simulation errors, the median ratio of between to total sum of squares (BSS/TSS) was computed for each block size. The resulting BSS/TSS correlated almost perfectly with the blocks sizes ($r = 0.94, 0.99$ for overlapping and non-overlapping scenarios, respectively), excluding thus this hypothesis. A literature review on clustering evaluation indices revealed that most of the existing measures, ARI included, are sensitive to class size unbalance [39]. Given that the noise cluster makes up from 0–75% of the simulations, we can suspect that such unbalance between signal and noise clusters is responsible for the observed behavior. The fact that this trend does not occur at low SNR, however, questions this explanation or suggests that the unbalance effect has a smaller impact at low SNR.

The hierarchy among methods is similar between blocks sizes, SNR levels and structures. iCluster and moCluster show the highest ARIs when $n_b < 20$ and remain high at $n_b = 20$. JIVE displays a sensitivity close to these two methods for all blocks sizes. iNMF and BCC, on the other hand, are the least sensitive with, however, a sharp improvement of iNMF for $n_b = 20$. Lastly, MDI did not return any results for $n_b < 20$, which again supports its lack of robustness against perturbations. Although the results with $n_b = 20$ are consistent with those obtained in Methods overview, they are not exactly equal since the numbers of variables (see Supplementary Materials) and repetitions by block size are slightly different.

Application: TCGA breast cancer data set

We now examine how the six methods compare on the TCGA breast cancer data set [2], the TCGA data being extensively used in the evaluation of integrative approaches [24, 26, 28, 30, 35]. The breast cancer data set consists of SNP, RNA, miRNA, DNA methylation and protein (reverse phase protein array) measurements in 825 patients. Here, the analysis is based on a subset of 348 patients assayed across all platforms, for which data were imputed and pre-processed by the authors of BCC (see BayesCC R package). Of note, SNP data were left aside by the authors. Because cancers are heterogeneous diseases, the diagnostic accuracy is essential for both the prognostic and the choice of treatment. The American Cancer Society classifies breast cancer into four molecular subtypes, HER2 enriched, basal (triple negative) and luminal A and B, based on the expression of proliferating protein Ki67 and the receptor status for estrogen (ER), progesterone (PR) and human epidermal growth factor 2 (HER2), as described in Table 3. Because tumors with similar immunohistochemistry and clinicopathological profiles may have different behaviors, recent omic approaches have sought to refine this classification by identifying new molecular signatures [40]. However, given that no consensus has emerged yet, we will consider the classification from the American Cancer Society (used in clinics) as gold standard and evaluate the integrative methods based on their consistency with subtypes derived from the receptor status.

Since Ki67 was missing in the data, luminal A and B subtypes could not be distinguished. For this reason, although integrative methods were run with 4 clusters on 348 patients, only the 84 patients annotated as basal and HER2 from the clinical data were kept in the computation of ARI. Similarly to the simulation studies, consensus clustering, GMM clustering on concatenated

Table 3. Breast cancer subtypes as defined by the American Cancer Society [41]

Subtype	Markers status
Basal	ER– PR– HER2–
HER2	ER– PR– HER2+
Luminal A	ER+ and/or PR+ HER2–
Luminal B	ER+ and/or PR+ HER2+ or High Ki67

matrices and GMM clustering on single omics were included in the comparison.

On Table 4, one can note that single-omic clusterings display a wide range of performances, suggesting that these omics are not impacted similarly during carcinogenesis. However, it cannot be excluded that this result is due to the higher number of features measured in mRNA and DNA methylation. Unexpectedly, a null ARI was found with proteins, which can be attributed to the high cluster unbalance obtained with GMM in this omic. Although these observations contrast with the high concordance between protein and mRNA subtypes reported in the original study [2], this may arise from a difference of samples used, our analysis being based on patients assayed on all platforms.

In line with the simulation results, all integrative methods but iCluster and JIVE outperformed single-omic approaches. The hierarchy among methods slightly differs with this obtained in the simulations; although moCluster and iNMF remain the top performing methods, they are closely followed by MDI then non-integrative approaches and BCC. iCluster and JIVE, on the other hand, present ARI 0.1 to 0.17 smaller than the others.

We then manually assigned (colored) clusters to their most plausible subtype based on Table 3: clusters with small percentages for all receptors were classified as basal, whereas those with high percentages of ER and PR were assigned to luminal A/B. As indicated above, luminal A and B clusters were merged due to the absence of Ki67 in the data. All methods successfully identified one basal and one to three luminal clusters, but none recovered the HER2 subtype. The absence of HER2 cluster and the over-representation of luminal ones are probably due to their subtype prevalence, larger in the former [41]. Six methods identified another cluster with average percentage values for all receptors; although matching no subtype, this cluster is most likely a mixture of HER2 and luminal patients. Because the eight approaches identified the four expected subtypes with a comparable, moderate accuracy, this step did not allow to further refine the method hierarchy.

In the same way as the simulations, this application confirmed that integrative approaches have an improved ability to identify common structures over single omics. Although moCluster and iNMF came first, Bayesian approaches surpassed iCluster and JIVE, in contrast with the results obtained in the simulations. We can, however, suppose that the use of sparsity could significantly improve the results of the latter.

Discussion and conclusion

Six popular integrative clustering methods, representative of matrix factorization and Bayesian approaches, were compared on simulations based on their (i) sensitivity and their ability to recover the (ii) number of clusters, (iii) common and specific structures across three data sets. Different simulation scenarios based on 12 combinations of models, SNR, data set dimension (iNMF-high-dimensional, sensitivity study) and overlaps between common and specific structures were tested to unveil methods' strengths and limitations.

The results from the simulations and application revealed that matrix factorization methods were on average better at identifying both common structures and the correct number of clusters; iCluster and moCluster outperformed the other methods on all criteria except on the application (iCluster) or the enumeration of the number of clusters (moCluster). Despite a probable lack of sensitivity, iNMF also showed a great ability to detect common clusterings and offered a homogeneity parameter, allowing the user to finely tune the

Table 4. Cluster profiles in terms of receptor percentages; consistency (ARI) between cancer subtypes and estimated clusterings. Clusters are colored according to their similarity to the 4 subtypes defined Table 3

Method	%ER	%PR	%HER2	ARI
Consensus clustering	97	89	11	0.52
	66	45	38	
	11	5	2	
	96	80	16	
Concatenation	97	89	7	0.52
	98	77	22	
	13	6	2	
	63	44	41	
iCluster	95	71	26	0.42
	96	85	8	
	90	81	18	
	12	5	9	
moCluster	13	6	2	0.57
	98	83	8	
	64	40	56	
	96	89	9	
iNMF	8	56	41	0.56
	97	9	6	
	13	6	7	
	100	87	8	
JIVE	96	84	9	0.40
	12	4	7	
	99	84	6	
	79	63	39	
BCC	70	49	43	0.51
	18	9	4	
	97	84	10	
	98	89	9	
MDI	94	87	10	0.55
	14	8	3	
	100	94	19	
	99	83	10	
mRNAs		–		0.50
DNA methylations		–		0.41
miRNAs		–		0.30
Proteins		–		0.00

matrix factorization between shared and specific structures, as depicted in Supplementary Figure 3. JIVE was generally close to the other matrix factorization methods, with however lower performances on the application and the detection of specific clusters. While BCC revealed a good ability to identify common structures, except on the iNMF-non-overlap and sensitivity scenarios, its main strength resides in its capability to simultaneously detect shared and specific structures. Lastly, MDI showed good performances in high SNR simulations and the application but had little robustness against data perturbations (noise and overlap between shared and specific structures). Of note, despite their longer runtime, the two Bayesian approaches were easier to parametrize. Additionally, we showed that neither the dimensionality nor the unbalanced number of features across data sets had an impact on the results. Some limitations of our work must be acknowledged. First, the evaluation criterion utilized throughout this work was the coherence between known and estimated sample clustering. A similar evaluation could also be performed at the variable level. Second, because a fair amount of time was invested in parameter tuning, we decided not to include feature selection in it. It would however be worth investigating the effect of penalization on the method performances. Third, although we highlighted pros and cons of these six methods through various simulation scenarios, method robustness could also have been evaluated by adding noise variables in varying proportions.

In addition to the presented benchmarking, our work demonstrated on all simulations the advantage of integrative methods over non-integrative ones in the identification of common structure, supporting their use in the identification of complex structures across omic layers.

Key Points

- The integration of multiple omics shows a clear improvement in clustering performance as compared to non-integrative methods.
- Matrix factorization methods are on average better at identifying common structure.
- Although iNMF showed a lack of sensitivity, it can finely be tuned to recover either common or specific structures.
- Despite moderate performances on shared clusters, BCC displayed the best ability to recover both structures simultaneously.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgments

We thank Vivian Viallon for helpful discussions. We also thank the IN2P3 Computing Center (Centre National de la Recherche Scientifique, Lyon-Villeurbanne, France) for providing high performing infrastructure.

Funding

BIOASTER investment funding (ANR-10-AIRT-03).

References

1. Sotiriou C, Neo SY, McShane LM, *et al*. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc Natl Acad Sci U S A* 2003; **100**(18):10393–8.
2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors. *Nature* 2012;**490**(7418):61.
3. Quigley DA, Fiorito E, Nord S, *et al*. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. *Mol Oncol* 2014;**8**(2):273–84.
4. Meng C, Kuster B, Culhane AC, *et al*. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics* 2014;**15**(1):162.
5. Ritchie MD, Holzinger ER, Li R, *et al*. Methods of integrating data to uncover genotype–phenotype interactions. *Nat Rev Genet* 2015;**16**(2):85–97.
6. Wirapati P, Sotiriou C, Kunkel S, *et al*. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008;**10**(4):R65.
7. Cavill R, Sidhu JK, Kilarski W, *et al*. A combined metabonomic and transcriptomic approach to investigate metabolism during development in the chick chorioallantoic membrane. *J Proteome Res* 2010;**9**(6):3126–34.
8. Cavill R, Jennen D, Kleinjans J, *et al*. Transcriptomic and metabolomic data integration. *Brief Bioinform* 2016;**17**(5): 891–901.
9. Ahmad A, Fröhlich H. Integrating heterogeneous omics data via statistical inference and learning techniques. *Genom Comput Biol*2016;**2**(1):e32.
10. Boulesteix AL, De Bin R, Jiang X, *et al*. IPF-LASSO: Integrative-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computat Math Methods Med* 2017; ID 7691937.
11. Sun H, Wang H, Zhu R, *et al*. iPEAP: integrating multiple omics and genetic data for pathway enrichment analysis. *Bioinformatics* 2014;**30**(5):737–9.
12. Kamburov A, Cavill R, Ebbels TM, *et al*. Integrated pathway-level analysis of transcriptomics and metabolomics data with IMPaLA. *Bioinformatics* 2011;**27**(20):2917–8.
13. Chalise P, Koestler DC, Bimali M, *et al*. Integrative clustering methods for high-dimensional molecular data. *Transl Cancer Res* 2014;**3**(3):202.
14. Wei Y. Integrative analyses of cancer data: a review from a statistical perspective. *Cancer Inform* 2015;**14**(Suppl 2):173–81.
15. Meng C, Zeleznik OA, Thallinger GG, *et al*. Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief Bioinform* 2016;**17**(4):628–41.
16. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. *Brief Bioinform* 2018;**19**(2):325–40.
17. Bersanelli M, Mosca E, Remondini D, *et al*. Methods for the integration of multi-omics data: mathematical aspects. *BMC Bioinformatics* 2016;**17**(Suppl 2):167–77.
18. Tini G, Marchetti L, Priami C, *et al*. Multi-omics integration—a comparison of unsupervised clustering methodologies. *Brief Bioinform* 2017;**167**.
19. Wang W, Baladandayuthapani V, Morris JS, *et al*. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. *Bioinformatics* 2013;**29**(2):149–59.
20. Yuan Y, Savage RS, Markowetz F. Patient-specific data fusion defines prognostic cancer subtypes. *PLoS Comput Biol* 2011;**7**(10):1–12.

21. Jennings EM, Morris JS, Carroll RJ, *et al*. Bayesian methods for expression-based integration of various types of genomics data. *EURASIP J Bioinform Syst Biol* 2013;**2013**(1):13.
22. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014;**11**(3):333–7.
23. Mosca E, Milanesi L. Network-based analysis of omics with multi-objective optimization. *Mol Biosyst* 2013;**9**(12):2971–80.
24. Shen R, Wang S, Mo Q. Sparse integrative clustering of multiple omics data sets. *Annals Appl Stat* 2013;**7**(1):269.
25. Shen R, Olshen AB, Ladanyi M. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* 2009;**25**(22):2906–12.
26. Meng C, Helm D, Frejno M, *et al*. moCluster: identifying joint patterns across multiple omics data sets. *J Proteome Res* 2016;**15**(3):755–65.
27. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. *J Roy Stat Soc B* 2001;**63**(2):411–23.
28. Lock EF, Hoadley KA, Marron JS, *et al*. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat* 2013;**7**(1):523–42.
29. Trygg J, Wold S. O2-PLS, a two-block (X-Y) latent variable regression (LVR) method with an integral OSC filter. *J Chemom* 2003;**17**(1):53–64.
30. Yang Z, Michailidis G. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics* 2016;**32**(1):1–8.
31. Zhang S, Liu CC, Li W, *et al*. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res* 2012;**40**:9379–9391.
32. Devarajan K. Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Comput Biol* 2008;**4**(7):e1000029.
33. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**(6755):788–91.
34. Kirk P, Griffin JE, Savage RS, *et al*. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics* 2012;**28**(24):3290–7.
35. Lock EF, Dunson DB. Bayesian consensus clustering. *Bioinformatics* 2013;**29**(20):2610–6.
36. Rand WM. Objective criteria for the evaluation of clustering methods. *J AmStat Assoc* 1971;**66**(336):846–50.
37. Hubert L, Arabie P. Comparing partitions. *J Classif* 1985;**2**(1):193–218.
38. Fritsch A, Ickstadt K. *et al*. Improved criteria for clustering based on the posterior similarity matrix. *Bayesian Anal* 2009;**4**(2):367–91.
39. de Souto MCP, Coelho ALV, Faceli K, *et al*. A comparison of external clustering evaluation indices in the context of imbalanced data sets. In: *Proceedings of Brazilian Symposium on Neural Networks, 2012*, 49–54. IEEE, Brazil.
40. Dai X, Li T, Bai Z, *et al*. Breast cancer intrinsic subtype classification, clinical use and future trends. *Am J Cancer Res* 2015;**5**(10):2929–43.
41. American Cancer Society. Breast Cancer Facts and Figures 2017–2018. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/breast-cancer-facts-and-figures/breast-cancer-facts-and-figures-2017-2018.pdf, Access date: 2017.

Resource

Pan-cancer multi-omics analysis and orthogonal experimental assessment of epigenetic driver genes

Andrea Halaburkova, Vincent Cahais, Alexei Novoloaca, Mariana Gomes da Silva Araujo, Rita Khoueiry,[1] Akram Ghantous,[1] and Zdenko Herceg[1]

Epigenetics Group, International Agency for Research on Cancer (IARC), 69008 Lyon, France

The recent identification of recurrently mutated epigenetic regulator genes (ERGs) supports their critical role in tumorigenesis. We conducted a pan-cancer analysis integrating (epi)genome, transcriptome, and DNA methylome alterations in a curated list of 426 ERGs across 33 cancer types, comprising 10,845 tumor and 730 normal tissues. We found that, in addition to mutations, copy number alterations in ERGs were more frequent than previously anticipated and tightly linked to expression aberrations. Novel bioinformatics approaches, integrating the strengths of various driver prediction and multi-omics algorithms, and an orthogonal in vitro screen (CRISPR-Cas9) targeting all ERGs revealed genes with driver roles within and across malignancies and shared driver mechanisms operating across multiple cancer types and hallmarks. This is the largest and most comprehensive analysis thus far; it is also the first experimental effort to specifically identify ERG drivers (epidrivers) and characterize their deregulation and functional impact in oncogenic processes.

[Supplemental material is available for this article.]

Although it has long been known that human cancers harbor both genetic and epigenetic changes, with an intricate interplay between the two mechanisms underpinning the hallmarks of cancer (Hanahan and Weinberg 2011), it is only with the fruition of large-scale international sequencing efforts that major enigmas of the cancer (epi)genome have started to be solved (Jones et al. 2016; Ng et al. 2018). One of the most remarkable findings of the international high-resolution cancer genome sequencing efforts, spearheaded by The Cancer Genome Atlas (TCGA), is the high frequency of genetic alterations in the genes encoding proteins that directly regulate the epigenome (referred to here as epigenetic regulator genes [ERGs]) (Gonzalez-Perez et al. 2013; Plass et al. 2013; Shen and Laird 2013; Timp and Feinberg 2013; Vogelstein et al. 2013; Yang et al. 2015). This high rate of ERG genetic deregulation constitutes a "genetic smoking gun," indicating that epigenetic mechanisms lie at the very heart of cancer biology. These discoveries have sparked a debate on the role of ERG deregulation (either through mutational or nongenetic events) in ERG expression and in the mechanisms underlying tumorigenesis and epigenome alterations that are rampant in virtually all human malignancies (Plass et al. 2013; Timp and Feinberg 2013). We also still lack a systematic understanding of the functional importance of ERG disruption in tumor development and progression, as well as its impact on cancer cell phenotype.

ERGs are a group of more than 400 coding genes in the human genome, most of which encode enzymes that add ("writers"), modify/revert ("editors"), or recognize ("readers") epigenetic modifications (Plass et al. 2013; Vogelstein et al. 2013) controlling a range of critical cellular processes. Based on the observation that many ERGs are frequently disrupted across different malignancies, they are candidates to be drivers of cancer development and progression, potentially acting as oncogenes or tumor suppressors (Plass et al. 2013; Vogelstein et al. 2013). Although several distinct definitions of "driver gene" exist in the literature (Sawan et al. 2008; Vogelstein et al. 2013), we define "driver genes" as those genes that, when deregulated (through somatic mutations, copy number variations, or aberrant expression), assume primary importance in tumor development such as conferring a selective growth advantage, immortalization, and invasiveness. This definition relies on inference models for driver prediction and functional data (based on the impact of the gene on cellular processes) compared to other methods that are mostly based on statistical models (largely driven by the mutation frequency of a gene) (Parmigiani et al. 2009; Meyerson et al. 2010; Lahouel et al. 2020). In line with this physiological definition, we refer to those ERGs that make a net contribution to tumorigenesis as "epigenetic driver genes" (henceforth called "epidrivers"). Our definition is different from that used by other investigators (Vogelstein et al. 2013), who define epidrivers as the genes (not necessarily among ERGs) that are aberrantly expressed through changes in DNA methylation and chromatin modifications and confer a selective growth advantage.

The products of ERGs are involved in processes such as DNA methylation, histone modification, chromatin remodeling, and other chromatin-based modifications, and many ERGs may have both histone and nonhistone substrates. All of these processes, in turn, are involved in the proper control of not only gene expression programs, required for the establishment and maintenance of cell identity and function, but also DNA repair, recombination, and genome integrity (Murr et al. 2006; Bell et al. 2011). Because common cancers represent the final outcome of a multistep process, epidriver-based disruption of cellular processes may not only assume a primary role at different stages of tumorigenesis but also constitute critical mechanisms underpinning cancer cell plasticity and

[1]These authors contributed equally to this work.
Corresponding author: hercegz@iarc.fr
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.268292.120. Freely available online through the *Genome Research* Open Access option.

emergence of cancer resilience. Here, we conducted a systematic and comprehensive pan-cancer investigation of (epi)genetics- and transcriptome-based deregulation of all ERGs using in silico data curation in clinical samples and characterization of the driver potential by different computational tools. We also developed and tested a conceptual framework for experimental identification and functional characterization of the mechanistically important epidrivers that reshape the epigenome and contribute to cancer phenotypes. This framework builds on the latest knowledge of the cancer (epi)genome and genomic databases and includes powerful new experimental models including state-of-the-art genome-editing screens, phenotyping, and functional genomics.

Results

A four-stage strategy was used to identify and characterize ERGs with cancer driver potential (Fig. 1A). We assembled a comprehensive compendium of ERG genes by literature mining and manual curation, resulting in a list of 426 genes coding for histone modifiers, DNA methylation regulators, chromatin remodelers, helicases, and other epigenetic entities (Fig. 1B; Supplemental Tables S1, S2). To identify the candidate epidrivers across different cancer types, we first used comprehensive in silico data mining of genetic and RNA expression alterations of ERGs using data from TCGA (Fig. 1A). Collectively, the data encompassed 33 different cancer types from 25 tissue types, with sequencing information from 10,845 tumor samples and 730 normal tissues, including a total of 90,144,805 genetic alterations, which encompassed 1,500,358 somatic mutations (single-nucleotide alterations [SNAs]) and 88,644,447 somatic copy number alterations (CNAs) (of which 4,294,698 were deep deletions/amplifications). We subsequently characterized, across various cancer types, the driver potential of ERGs based on ConsensusDriver scores (Bertrand et al. 2018), which we complemented with our proposed Pan-Cancer Driver and Multi-Omics Driver scores, and the implications of these ERGs in cancer hallmark pathways. Finally, we performed an orthogonal validation of the driver potential of 426 ERGs in cell line models in comparison with the findings from the clinical samples (Fig. 1A).

Pan-cancer analysis of genetic alterations in ERGs

To identify potential epidrivers, we first analyzed the frequency of genetic disruption of ERGs (vs. all genes) across malignancies from different anatomical sites. Our analysis revealed that the predominant genetic alterations in ERGs were deep amplifications or SNAs, depending on the cancer type (Fig. 2A,B). In comparison, the predominant genetic alteration in all human genes was deep amplification in most cancer types (Supplemental Fig. S1A,B). Overall, higher proportions of amplifications than deletions were observed in ERGs or in all genes across all cancer types except a few (mainly, DLBC and PRAD) (Fig. 2A,B; Supplemental Fig. S1A–C). Some cancer types (e.g., OV) had predominately CNAs with almost no SNAs (Fig. 2A–D; Supplemental Fig. S1A,B). In most cancer types, the CNAs (Supplemental Fig. S1D) and SNAs (Supplemental Fig. S2) were uniformly distributed across chromosomes, with the exception of GBM, KIRP, and UVM, which showed CNAs in specific chromosomes (Supplemental Fig. S1D).

Many specific ERGs were identified as being genetically altered at noticeably high levels in different malignancies (Fig. 2E, F). In particular, SNAs in *IDH1* (Fig. 2E,G; Supplemental Fig. S2) and deep CNAs in *ACTL6A* (Fig. 2F,H) had high proportions of alterations, exceeding 40% of samples in LGG and LUSC, respectively. Several ERGs had the highest mutation frequency repeatedly in many cancer types, namely the *KMT2C/D* family (seven cancers), *ARID1A* (five cancers), *BAP1* (three cancers), and *ATRX* (three cancers) (Fig. 2E,G; Supplemental Table S3). A similar observation was made for deep CNAs in ERGs, namely *BOP1* (four cancers), *ATAD2* (four cancers), *MECOM* (three cancers), and *PHF20L1* (three cancers) (Fig. 2F,H). A larger percentage of ERG alterations was also observed when both deep and shallow CNAs were included (Fig. 2C, D; Supplemental Fig. S1C,D). Among the top ERGs altered by deep CNAs, the majority showed amplifications, with the exception of *HR*, *PHF11*, and *SETB2*, which were commonly deleted in many cancer types (Fig. 2H). Frequently amplified ERGs often co-occurred in the same tumor sample in many cancer types; in particular, the aforementioned pan-cancer recurrent genes *BOP1*, *ATAD2*, and *PHF20L1* highly co-occurred (Supplemental Fig. S3A). These co-occurrences remained prominent even when the analysis was focused only on deep amplifications/deletions across tumors (Supplemental Fig. S3B,C). Moreover, the family of TDRs (*TDRKH*, *TDRD10*, and *TDRD5*) highly co-occurred together. Generally little overlap was observed between the genes with a high frequency of SNAs and those with a high frequency of CNAs (except for a few ERGs) (Fig. 2E–H).

When ERGs were stratified by functional groups, similar total proportions of genetic alterations were seen among ERG classes (Fig. 2I). DNA methylation writers and editors were characterized by a prominent proportion of SNAs in several cancer types, compared with other ERG classes (Fig. 2I; Supplemental Fig. S4A,B). Indeed, DNA methylation modulators appeared among the top SNA profiles (Fig. 2G) but not among the top CNA profiles (Fig. 2H) of ERGs. Moreover, in many cancer types, DNA methylation writers or editors, which are among the smallest ERG classes, were the group showing the largest percentage of genetically altered ERGs (Supplemental Fig. S4A,B). Among ERGs that could be classified as tumor suppressors, *KMT2D*, *KMT2C*, *ARID1A*, *ATRX*, *CREBBP*, and *PBRM1* were frequently mutated in many cancer types, whereas oncogenic ERGs were each mutated in specific cancer types, mainly *IDH1* in LGG and *DNMT3A* in LAML (Supplemental Fig. S5A,B).

Pan-cancer analysis of RNA expression in relation to genetic and DNA methylome aberrations of ERGs

The second approach in our analysis focused on RNA expression deregulation (by RNA-seq) of ERGs across cancer types. For each cancer type, the analysis consisted of two parts: expression variation across tumor samples relative to one another (independent of the corresponding normal tissue) and expression changes in tumor relative to adjacent normal tissue. By integrating genetic and transcriptomic information matched to the same samples (using the TCGA database) a higher proportion of tumor samples showed significantly increased ERG expression (Z-score > 2) relative to down-regulation (Z-score < −2) (Fig. 3A), in line with the observed higher proportion of samples with ERG amplifications than with deletions (Figs. 2A, 3A).

Amplifications and deletions significantly correlated (false discovery rate [FDR] < 0.05) positively with increased and decreased expression, respectively, in all cancer types and chromosomes (Fig. 3A,B), except for Chromosome X, because of a statistical artifact (Methods; Supplemental Fig. S6A,B). SNAs significantly correlated (FDR < 0.05) negatively or positively with expression across tumor samples and chromosomes, so the correlation was not

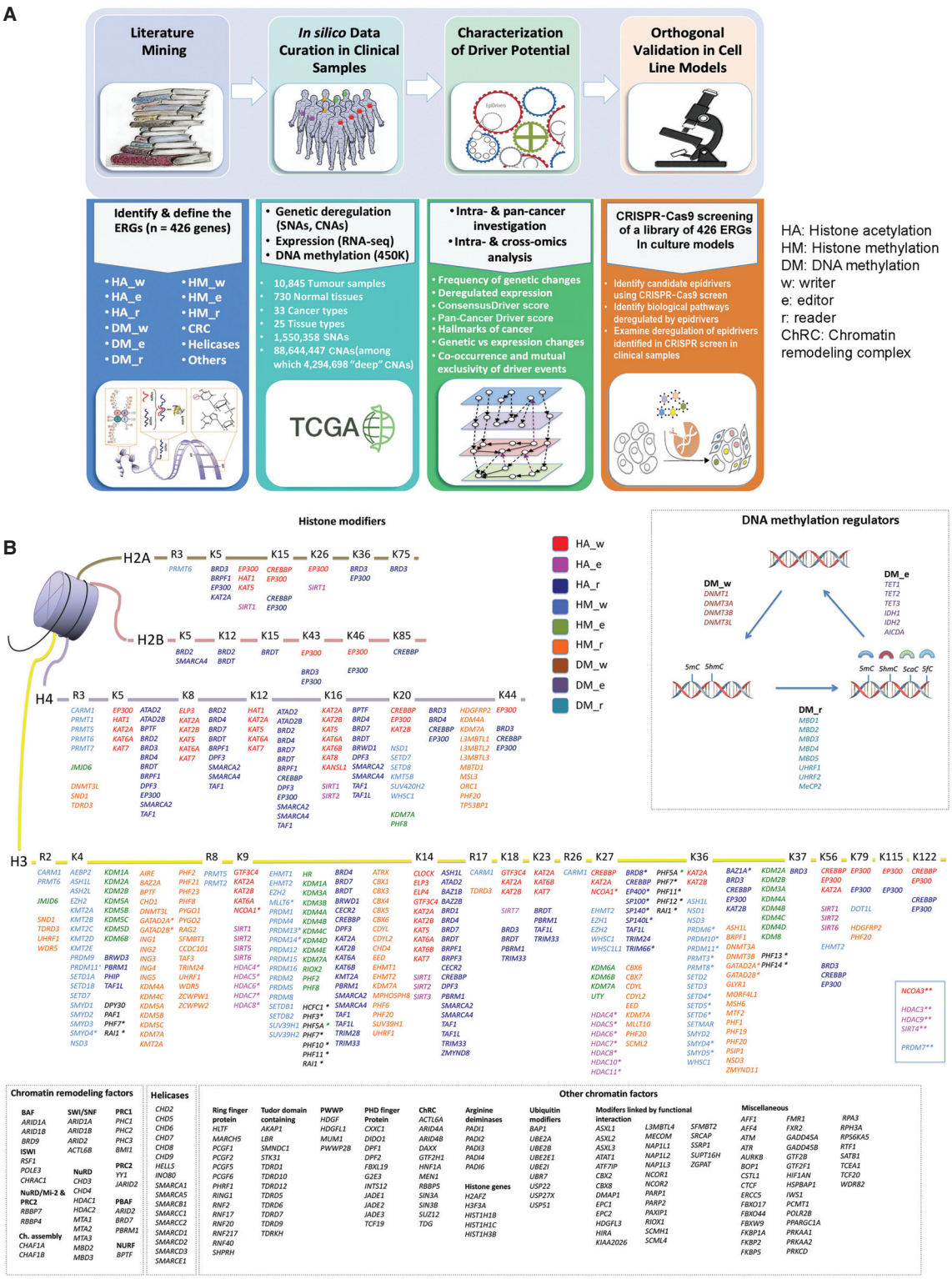


Figure 1. Study design. (*A*) A four-stage approach to identify and characterize ERGs with cancer driver potential. (*B*) The compendium of ERGs curated and analyzed, comprising 426 genes classified into histone modifiers, DNA methylation regulators, chromatin remodeling factors (ChRC), helicases, and other chromatin modifiers (some of which were further divided into subgroups based on function or their presence in molecular complexes). Histone acetylation, histone methylation, and DNA methylation modifiers are further stratified each into "writers" (w), "editors" (e), and "readers" (r). (*) The histone modifying genes whose functions are not well characterized and which were, therefore, assigned based on ENCODE ChIP sequencing data; (**) the histone modifying genes without assignment of residues in the histone tails.

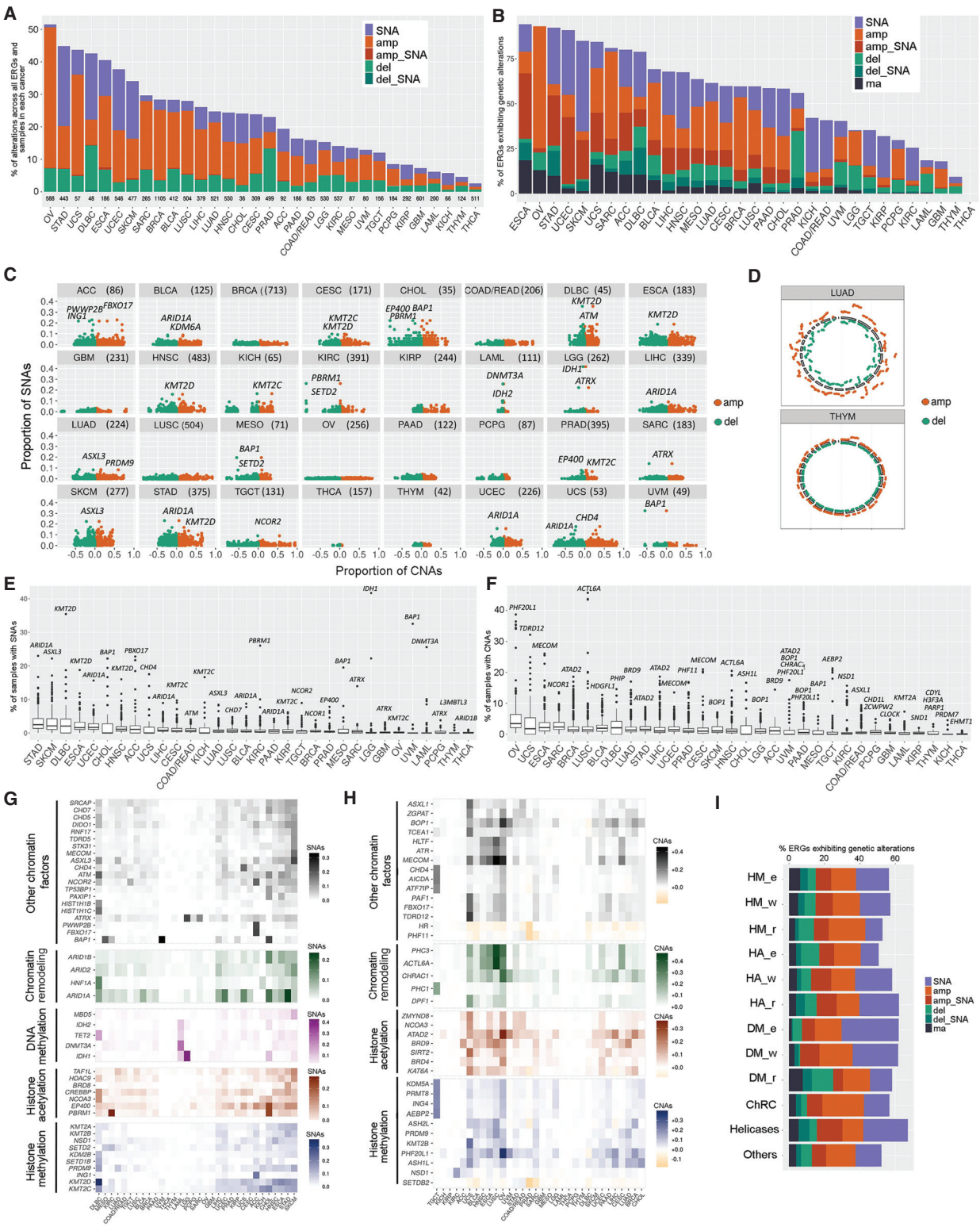


Figure 2. Pan-cancer analysis of genetic alterations across ERG categories and classes. (*A*,*B*) The percentage of samples with genetic deregulation in ERGs (*A*) and the percentage of ERGs showing different types of genetic deregulation (*B*), by cancer type. ERGs are considered altered if at least 1% of samples harbor these genetic aberrations. (*C*) Proportion of samples with SNAs versus that with deletions (−1, −2) or amplifications (+1, +2) in ERGs for each cancer type. Each gene is represented by two dots (red and green) depicting amplified and deleted CNAs, respectively. (*D*) Circos plots showing the relative amount of deregulation in CNAs by chromosomal distribution in two representative cancer types (LUAD and THYM, characterized by high and low CNA burden, respectively). The level of CNAs for each ERG was calculated as the proportion of samples considering all types of CNAs (amplification = +1, +2 and deletion = −1, −2) in ERGs in each cancer type. (*E*,*F*) Box plots showing the percentage of samples with SNAs (*E*) and deep CNAs (*F*) by gene and by cancer type. The most deregulated ERGs are highlighted for each cancer type. (*G*,*H*) Heatmaps representing the top genetically deregulated genes showing SNAs (*G*) and CNAs (*H*) in at least 10% and 15%, respectively, of the samples for any cancer type. Only samples with deep CNAs were included. ERGs are grouped into functional categories as indicated. (*I*) The percentages of ERGs that show genetic alteration among all cancer types by functional groups. Genetic alterations: (SNA) single-nucleotide alteration, (amp) deep copy number amplification, (amp_SNA) deep amplification co-occurring with SNA, (del) deep copy number deletion, (del_SNA) deep deletion co-occurring with SNA, and (ma) multiple alterations. In cases in which both types of CNAs (amplification and deletion) of one gene were present in the samples, we reported in *B* and *H* the alteration that was at least twice as prevalent as the other; otherwise, the alteration was reported under the multiple alteration category.

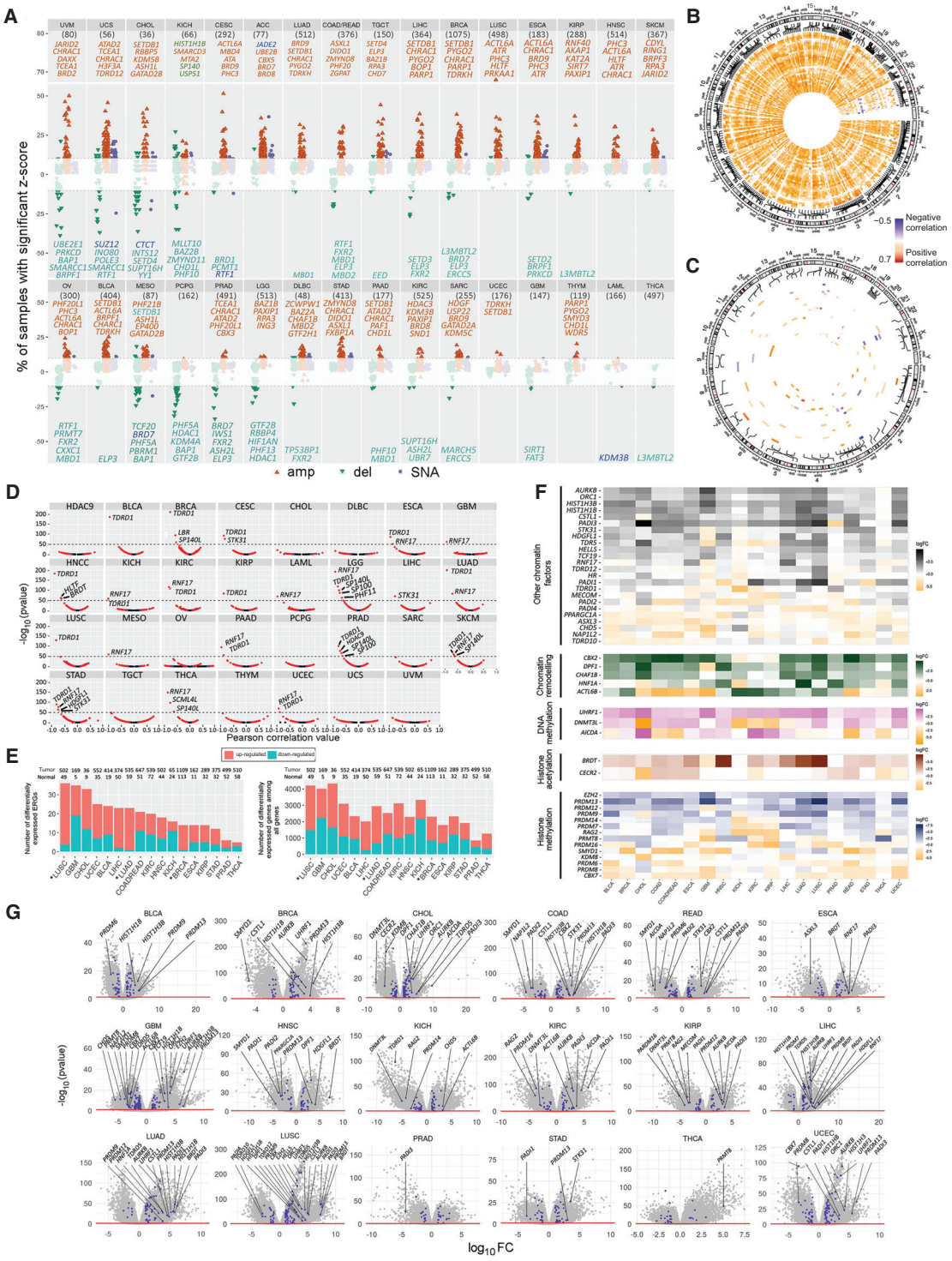


Figure 3. RNA expression alterations of ERGs across cancer types, in relation to genetic and DNA methylome variations. (*A*) Multi-omics plot of SNA, CNA, and RNA expression alterations across ERGs and cancer types. Amplifications, deletions, and SNAs were annotated as described in Methods. The most deregulated ERGs in RNA expression (with the *y*-axis value above 10) are highlighted for each cancer type. (*B*,*C*) Circos plots showing Pearson's correlation between CNAs (*B*) or SNAs (*C*) and expression *Z*-scores in different cancer types across the chromosomal regions. Positive and negative correlations are indicated in orange and blue, respectively. Only ERGs with correlation (R^2) > 30% and FDR < 0.05 in at least in one cancer type were considered for the analysis in *B*; the R^2 limit was set to 10% in *C*. (*D*) Expression quantitative trait methylation (eQTMs) analysis showing Pearson correlation values (*x*-axis) between RNA (RSEM counts) and methylation (beta) levels of promoter CpGs for each ERG in different cancer types. The line bar indicates highly significant CpGs [−log(*P*-value) > 50]. Red, blue, and black dots represent CpGs with FDR < 0.05, $P < 0.05$, and $P > 0.05$, respectively. (*E*) Number of ERGs or all genes with differential RNA expression in tumor relative to adjacent normal tissues for each cancer type (|log FC| > 2 and FDR < 0.05). The star denotes a *P*-value < 0.05 by a two-sample test of proportions of up- versus down-regulation. (*F*) Heatmaps showing the most differentially expressed ERGs comparing tumor samples with adjacent normal tissues among cancer types. Only the top differently expressed ERGs with |log FC| > 3 and FDR < 0.05 are annotated. (*G*) Volcano plots showing differentially expressed ERGs in tumors relative to adjacent normal tissues. ERGs are shown in blue (|log FC| > 1), and the most deregulated ERGs with |log FC| > 3 are highlighted for each cancer type (FDR < 0.05). Sample sizes for each cancer type are indicated in *A* and *E*.

unidirectional (Fig. 3A,C). ERG-specific analysis revealed some ERGs with noticeably high expression aberrations within tumor samples (Fig. 3A). Among them, several ERGs were repeatedly up- or down-regulated in several cancer types; the primary genes were *CHRAC1*, *PHC3*, *and BOP1*, which were among the top 10 most up-regulated ERGs recurrently observed in 18, 9, and 7, respectively, out of 33 cancer types (Fig. 3A, only the top 5 ERGs are shown; see Supplemental Table S3 for the top 10 ERGs). Many of the ERGs with the highest deregulated expression are the same ERGs with the highest CNAs of the corresponding cancer type (Fig. 3A vs. Fig. 2F).

Because epigenetic inactivation could be an additional mechanism for aberrant expression of ERGs, we performed concurrent analyses of transcriptomic and DNA methylome data available for the tumor samples in the TCGA. We correlated the methylation with RNA-seq levels by limiting the comparison to CpGs in promoter regions (−1000 to +500 bp of TSS) and RNA transcripts of the overlapping gene. The most significantly correlated CpG for each gene is shown in Figure 3D, and results for all analyzed CpGs are provided in Supplemental Table S4 and the figshare repository (https://figshare.com/articles/Supplemental_data/12613220). All CpGs that were highly significant [−log(*P*-value) > 50] were negatively correlated with the expression of their corresponding gene, and many of them were recurrent in several tumor types, namely CpGs in *TDRD1* ($n = 16$ tumor types), *RNF17* ($n = 13$ tumor types), *SP140L* ($n = 5$ tumor types), and *STK31* ($n = 3$ tumor types) (Fig. 3D).

Comparing the expression levels of ERGs in tumor relative to adjacent normal tissue in each cancer type also revealed a predominant pattern of overexpressed ERGs in most cancer types, except GBM and KICH (Fig. 3E). Similar observations were made when all human genes were analyzed (Fig. 3E). ERG-specific analysis revealed several ERGs with significant deregulation of expression (FDR < 0.05) (Fig. 3F,G) and recurrence in several cancer types. Several genes had similar recurrence across several cancer types, namely *PADI3* (in 11 of 18 cancers), *PRDM13* (in 10 of 18 cancers), *AURKB* (in 9 of 18 cancers), and *HIST1H1B* and *HIST1H3B* (each in 8 of 18 cancers), based on the selection of only the top genes (FDR < 0.05 and $\log_{10}$ fold change [$\log_{10}$FC] > 3) (Fig. 3G).

Characterizing the driver potential of deregulated ERGs across cancer types

The third strategy in our analysis to characterize the potential driver roles of ERGs was based on ConsensusDriver, a novel approach that provides a systematic way to integrate the strengths of various driver prediction algorithms (Bertrand et al. 2018). The ERGs with a potential driver role (ConsensusDriver score > 1.5) are shown for each cancer type (Fig. 4A) and are significantly enriched relative to the 233 total genes (Bailey et al. 2018) that have a driver score > 1.5 ($P = 4.0 \times 10^{-22}$, Fisher's exact test). Six additional ERGs would still be classified as drivers at a score < 1.5 but with manual curation by Bailey et al. (2018), and these are *ATR*, *EZH2*, *HIST1H1C*, *PHF6*, *SMARCB1*, and *TET2*. The ConsensusDriver score matched to a high extent with the driver potential predicted based on SNA frequencies in each cancer type, and to a lesser extent with that predicted based on CNA, FC, or *Z*-scores (Fig. 4A); the latter three, if matching with ConsensusDriver score, never occurred without SNAs, further emphasizing the importance of SNAs in the derivation of ConsensusDriver score (Fig. 4A). *IDH1* had the highest ConsensusDriver score, as evident in LGG, and this ERG showed a driver role in six other cancer types, which explains why it additionally had the highest pan-cancer ConsensusDriver score (PANCAN) (Fig. 4A).

In comparison, *ARID1A* was the ERG with the most frequent driver score, appearing in 13 cancer types, although with a relatively weak to modest driver role in individual cancer categories. ConsensusDriver ERGs were enriched in several gene families (namely *ARID(1A/2)*, *ASXL1/2*, *CHD3/4/8*, *IDH1/2*, *KDM(5C/6A)*, *KMT2A/B/C/D*, *NSD1/2*, and *SMARCA1/4*) as well as in UCEC ($n =$ 14 ERGs) and BLCA ($n = 10$ ERGs) (Fig. 4A). Genetic deregulation of ConsensusDriver ERGs often co-occurred in the same sample across several cancer types, with *KMT2D* and *ARID1A* having the highest co-occurrence scores. The ERGs within the same *KMT2A/B/C/D* family highly co-occurred together even though they were not mapped to the same chromosomes (Supplemental Fig. S3C), whereas *IDH1* and *IDH2* were mutually exclusive (Fig. 4B).

We complemented ConsensusDriver (weighted for SNAs) with our Multi-Omics Driver score that is weighted for each of SNAs, CNAs, and expression aberrations. The Multi-Omics Driver scores for all ERGs across cancer types is shown in Figure 4C and Supplemental Table S5 (the figshare repository: https://figshare.com/articles/Supplemental_data/12613220). This score revealed ERGs with a high Multi-Omics Driver score in most cancer types (such as ATAD2) against those showing single driver score (such as IDH1, which has a high SNA driver score in LGG but relatively low CNA and expression driver scores). We next formulated another score, the Pan-Cancer Driver score, that additionally weights for pan-cancer coverage on top of SNAs, CNAs, and expression aberrations (Fig. 4D). *ATAD2* had the highest Pan-Cancer Driver score, showing all the SNA, CNA, and expression *Z*-score alterations in many cancer types. When we considered the top 39 Pan-Cancer Driver genes, representing a sample size identical to that identified from ConsensusDriver, we found several driver ERGs to be similarly represented in both sets, namely, *SMARCA4* (score = 11), *ASXL1* (score = 12), *BAP1* (score = 26), *KMT2B* (score = 38), and *MECOM* (score = 37). HM and HA, but not DM, modulators were highly represented among the top 100 Pan-Cancer Driver ERGs, probably because DM modulators are mostly altered by SNAs (Fig. 2G vs. 2H), and hence are characterized by ConsensusDriver (e.g., *IDH1* and *DNMT3A*) (Fig. 4A) rather than Pan-Cancer Driver (Fig. 4D; Supplemental Fig. S5B) profiles. Similarly, *ARID1A*, which is characterized predominantly by the SNA type of genetic alterations, showed ConsensusDriver potential in many cancer types (Fig. 4A) but did not appear among the top 100 Pan-Cancer Driver ERGs (Fig. 4D; Supplemental Fig. S5B). Genetic deregulation of Pan-Cancer Driver ERGs often co-occurred in the same sample across several cancer types (Fig. 4E).

Next, we investigated whether epidrivers are enriched in pathways affecting the 10 hallmarks of cancer. Our compendium of 426 ERGs was significantly enriched in four hallmarks, namely genome instability and mutation, evading growth suppressors, sustaining proliferative signaling, and enabling replicative immortality (Fig. 4F; Supplemental Fig. S7A,B), further supporting a driver role of ERGs in tumorigenesis and characterizing the nature of biological pathways in which ERGs play a functional role in cancer. These four hallmarks were also topmost significant in the 39 ConsensusDriver ($P < 0.05$), top 39 Pan-Cancer Driver ($P < 0.1$), and 42 multi-omic driver ($P < 0.05$) ERGs (Supplemental Fig. 7C).

Orthogonal CRISPR-Cas9 screen to assess the driver potential of ERGs in epithelial-to-mesenchymal transition (EMT)

We conducted a CRISPR-Cas9 screen using a custom-made lentiviral CRISPR library consisting of 1649 gRNAs targeting all 426 ERGs (Fig. 5A; Supplemental Fig. S8) and A549 lung cancer cells stably

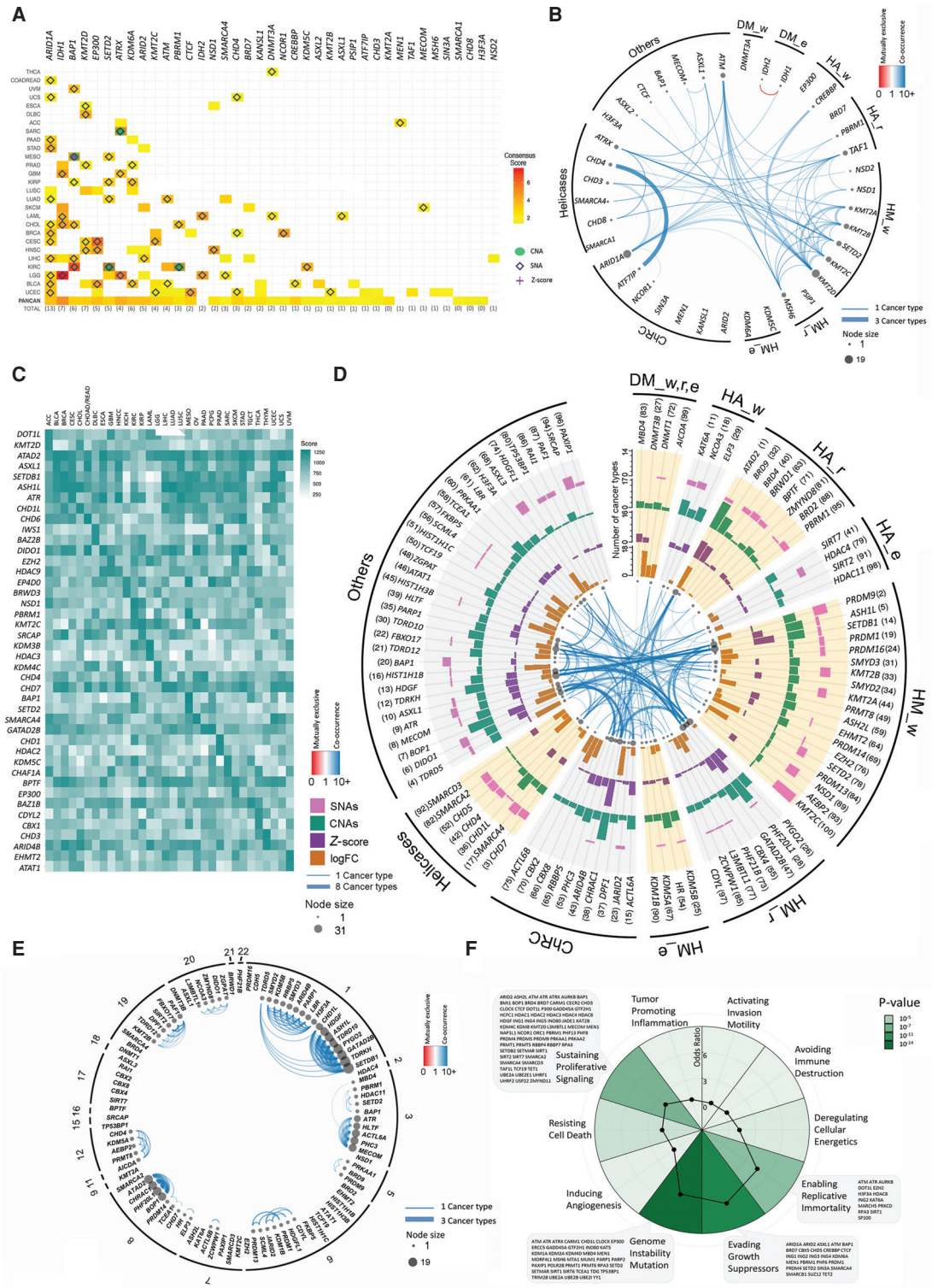


Figure 4. Characterization of ERG driver potentials. (*A*) Heatmap showing the ConsensusDriver scores (with values ranging from 1.5 to 7.5) as obtained by Bailey et al. (2018). ERGs with a score ≥ 1.5 in at least one cancer type are shown. The top 10 deep amplifications or deletions (green circles), SNAs (blue empty diamonds), or significant *Z*-score (purple crosses) of each cancer are overlapped onto the heatmap. (*B*) Significant (FDR < 0.05) co-occurrence and mutual exclusivity for ConsensusDriver ERGs in a pan-cancer analysis. The node size is proportional to both the number and thickness of its connections with other nodes. Blue and red edges represent co-occurrence (odds ratio [OR] > 1) and mutual exclusivity (OR < 1), respectively. The transparency of the edges indicates the average OR across cancer types, and their thickness is proportional to the number of cancer types in which the OR is significant. The co-occurrence filter was set to at least 5% of the samples per cancer type (Methods). (*C*) Heatmap of the Multi-Omics Driver scores of ERGs per cancer type. The ERGs shown represent a pooled set of the top three ERGs in each cancer type, as ranked by the mutli-omics driver score. (*D*) Top 100 ERGs by Pan-Cancer Driver score using SNA (5% of samples), CNA (5% of samples), and expression data (15% of samples with significant *Z*-score or FDR < 0,05 with $\log_{10}$FC > 1). Results are represented as bar plots counting the number of cancers in which a given gene has a particular genomic or expression alteration. From outer to inner track: (1, pink) SNAs; (2, green) CNAs; (3, purple) *Z*-score; (4, orange) $\log_{10}$FC. Inside the last track, co-occurrence or mutual exclusivity was calculated as in *B*, except that the co-occurrence filter was set to at least 10% of the samples per cancer type. Genes are aggregated by their functional features. (*E*) Significant co-occurrence for the top 100 ERGs by Pan-Cancer Driver score. Co-occurrence or mutual exclusivity was calculated as in *D* but ordered instead by chromosome number. (*F*) Spider pie chart showing enrichment of the 426 ERGs in pathways affecting the 10 hallmarks of cancer; the corresponding *P*-values and ORs are illustrated by green gradients and black spots, respectively. The names of ERGs overlapping with the four significantly enriched hallmarks are indicated.

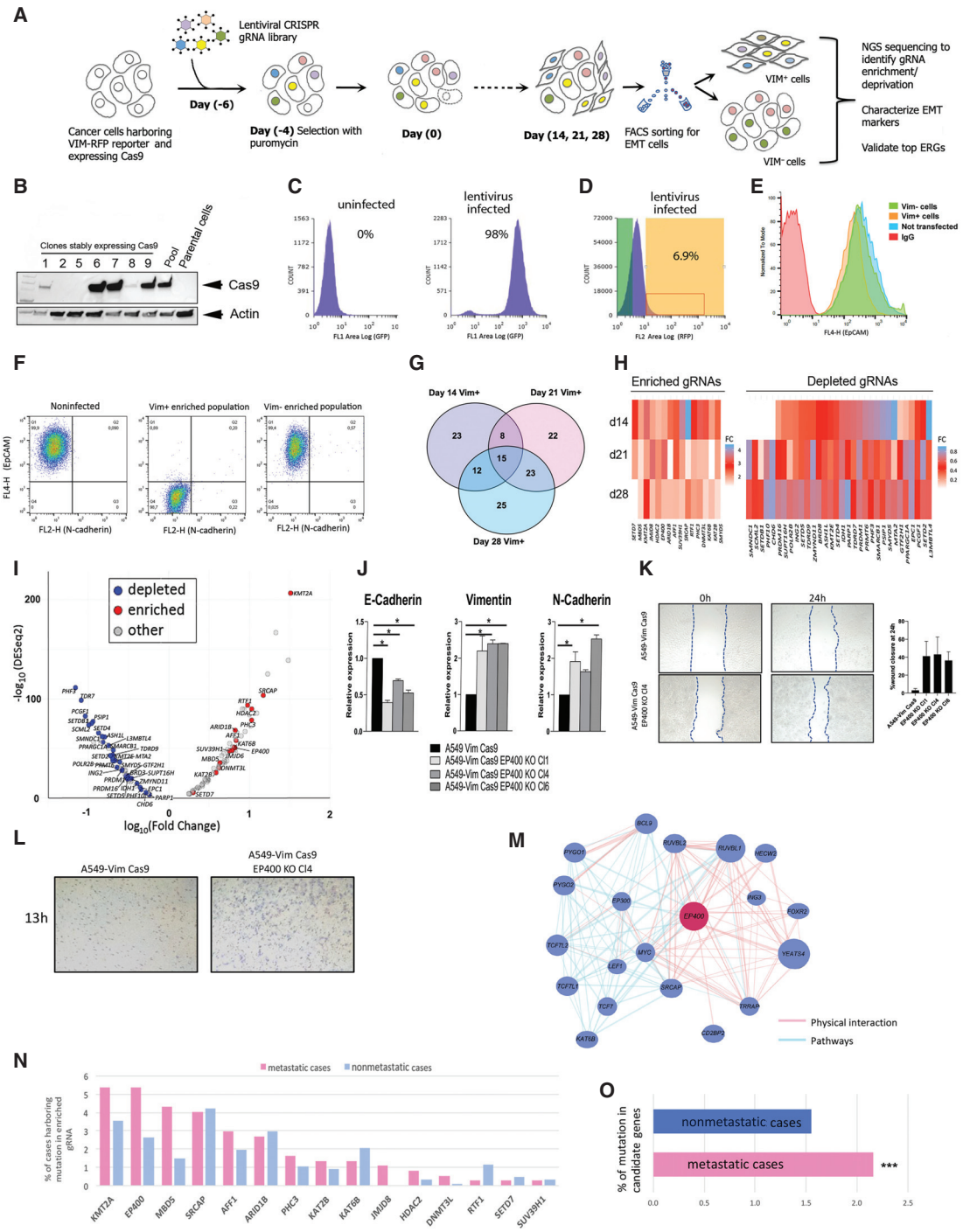


Figure 5. CRISPR-Cas9 screen to perform orthogonal assessment of the driver potential of ERGs in EMT. (*A*) The screening strategy used to identify positive and negative regulators of EMT among ERGs. (*B*) Western blot analysis of Cas9 expression in A549 lung cancer cells. "Pool" represents a heterogeneous population of transduced and stably Cas9 expressing cells derived from the parental cells. Individual cell clones derived by cloning rings are numbered 1, 2, 5, 6, 7, 8, and 9. Actin beta was used to normalize for equal loading. (*C*) Validation of the transduction efficiency of the lentiviral CRISPR ERG library 10 d after puromycin selection using FACS compared with uninfected A549 cells. (*D*) Enrichment of vimentin-positive (VIM^+) population analyzed by FACS after CRISPR ERG library transduction at day 14 after puromycin selection. (*E*) Validation of cell sorting for the enrichment of VIM^+ population by FACS based on the fluorescent antibody EPCAM (EPCAM loss is associated with the mesenchymal cell state) of VIM^+, vimentin-negative (VIM^-), uninfected cell line, and negative control antibody IgG. (*F*) Confirmation of cell enrichment for VIM^+ and VIM^- fractions after sorting. FACS-sorted VIM^+ and VIM^- populations were grown in culture for 2 wk and analyzed by FACS after staining with cadherin 2 (also known as N-cadherin) antibodies. (*G*) The overlap of the top EMT-associated ERG gRNAs after Illumina MiSeq deep sequencing; the numbers are derived from two statistical methods (DESeq2 and edgeR) at days 14, 21, and 28 after transduction. (*H*) Heatmap showing the top ERGs based on enriched and depleted gRNAs at days 14, 21, and 28 after transduction compared with day 0. (*I*) Volcano plot of ERG gRNAs at day 28 after transduction. (*J*) Expression analysis by qRT-PCR of EMT markers (cadherin 1 [also known as E-cadherin], vimentin, and cadherin 2) on single targeted A549-VimCas9 clones following *EP400* loss of function, relative to expression in the parental A549 Vim Cas9 cell line. (*) $P < 0.05$, indicates results of one-way ANOVA test. Error bars are SEM of $n = 2$. (*K*) Representative image of scratch assay performed on the parental cell line and three generated *EP400* KO clones at day 0 and after 24 h (*left*). On the *right*, a graph plot showing percentage area closure 24 h after the scratch as averaged of at least six areas analyzed for each clone and for the parental cell line. Experiments were performed in duplicates. Experiments were performed in duplicates. (*L*) Transwell migration assay showing increase of migration at 13 h for A549-Vim Cas9 *EP400* KO Cl4 compared to the parental cell line. (KO) Knockout; (Cl) clone; (vim) vimentin. (*M*) An example of network analysis of selected top ERGs (*EP400*) associated with the EMT population, obtained with the GeneMANIA package. (*N*,*O*) The bar plots show the mutation frequency of EMT-specific ERGs (identified in the CRISPR-Cas9 screen) in clinical samples from nonmetastatic (M0) and metastatic (M1) subsets (based on the annotation of TCGA samples).

expressing Cas9 (Fig. 5B,C), aimed at an orthogonal in vitro assessment of driver potential of ERGs. The screen was conducted in A549 lung cancer cells, considered a gold standard for studying the epithelial-to-mesenchymal transition (EMT), a cellular program conferring on cancer cells multiple traits associated with higher-grade malignancy, which may have an underlying epigenetic mechanism (Tam and Weinberg 2013). In the A549 cells, a red fluorescent protein (RFP)-tagged reporter is under the control of the endogenous vimentin promoter, thereby permitting the real-time monitoring of the transition from epithelial to mesenchymal status of the cells (activation of vimentin-RFP expression, mesenchymal marker) (Fig. 5D–F). The transduced A549-Vim Cas9 cells were further grown and collected at days 14, 21, and 28 after transduction, followed by flow cytometry (FACS) sorting to enrich for vimentin-positive (VIM^+) and vimentin-negative (VIM^-) fractions (Fig. 5C,D).

VIM^+ cells were further confirmed to express additional markers that are associated with the mesenchymal cell state, including cadherin 2^{high}, cadherin 1^{low}, and $EPCAM^{low}$ (Fig. 5E,F; Supplemental Fig. S9). These EMT markers remained stable after a prolonged culturing of VIM^+ cells (Fig. 5F), showing RFP fluorescence in A549 cells that provides a quantitative readout of EMT. To identify potential regulators of EMT among ERGs, we subjected the cells to deep sequencing followed by enrichment or depletion analysis of gRNA-targeting ERGs across the three different time points (days 14, 21, and 28) (Fig. 5G–I; Supplemental Fig. S10A–C). The top most significant hits identified belonged predominantly to histone writers, histone readers, and chromatin remodelers (Supplemental Table S6) and several biological pathways (Supplemental Fig. S11). Among the top most significant EMT-specific hits, one-third belonged to the category of histone methylation writers, whereas several ERGs (including *KMT2A*, *EP400*, *MBD5*, and *SRCAP*) were found to be frequently targeted by genetic alterations in several cancer types (Fig. 2E).

To validate our findings, we knocked out individually several of the identified targets in A549 cells by using three to four distinct gRNAs for each gene and analyzed the changes in expression of the epithelial marker cadherin 1 and the EMT markers cadherin 2 and vimentin in the mutant clones (Fig. 5J; Supplemental Fig. S12). Moreover, because EMT is associated with an increased tumor invasiveness, we evaluated whether an increase in the migration capacity was observed in cells upon loss of the candidate epidrivers of EMT by performing scratch and transwell migration assays. Indeed, loss of several of the epidrivers candidates led to a significant gain in EMT markers and was accompanied by a gain in the migration capacity of cells (Fig. 5K,L; Supplemental Fig. S12). This was significant in all knockout clones of *EP400*, *KAT2B*, *ARID1B* clone 2, and *MBD5* clone 10 (Supplemental Fig. S12).

Next, we assessed the potential link of EMT-specific, enriched, or depleted ERGs to biological pathways using different gene set enrichment bioinformatics tools and found that the *NOTCH1*, *WNT*, and *TP53* pathways were highly correlated with EMT-associated ERGs identified in the CRISPR-Cas9 screen (Supplemental Fig. S11A–C). We further applied the GeneMANIA prediction tool (Warde-Farley et al. 2010) to the top ERGs identified in our screen and found several directional dependencies (predominantly through physical interactions and common pathways) (Fig. 5M; Supplemental Figs. S11D, S13). *EP400*, *KMT2A*, *SRCAP*, and *KAT2B* were found to have direct interactions with several genes known to be involved in EMT, including *PYGO1*, *PYGO2*, and *TWIST1*, and a substantial number of genes known to form multiprotein complexes, thereby, connecting previously uncharacterized complexes/pathways to the EMT process. Finally, the top ERGs identified in our CRISPR-Cas9 screen (including *MBD5* and *JMJD8*) were significantly more frequently mutated in metastatic cancer cases compared with their nonmetastatic counterparts across 21 different cancer types (Fig. 5N,O), further corroborating the findings of the CRISPR-Cas9 screen that EMT-specific ERGs may be involved in conferring on cancer cells invasiveness and metastatic potential.

Identifying epidrivers involved in sustaining proliferation of cancer cells

To further expand our finding on the involvement of ERGs in hallmarks of cancer, we next applied the CRISPR library targeting all 426 ERGs on the A549-Vim and an independent cell line (MCF10A cells, the human mammary epithelial cell line widely used in vitro model for studying oncogenic transformation) and analyzed the ERGs involved in sustaining cell proliferative capacity. To this end, the cells A549-Vim and MCF10A cells expressing Cas9 were infected with the CRISPR library. Following selection, the cells were collected at different time points, subjected to deep sequencing, and analyzed for significantly enriched and depleted gRNAs (Fig. 6A; Supplemental Figs. S14, S15). We revealed 56 gRNAs that are enriched in MCF10A cells over the passages, 15 of which were also detected using DESeq2 (an independent statistical analysis method) (Supplemental Fig. S14B,C). The cell cycle was found to be among the top pathways enriched on the list of genes associated with enriched gRNAs (Supplemental Fig. S14D), consistent with the notion that the loss of function of the ERGs is linked with an increase in proliferation of MCF10A cells. KEGG analysis also revealed NOTCH and FOXO signaling pathways, two major pathways involved in breast cancer development, and several of the identified putative epidrivers (i.e., *ATRX*, *PHF11*, *NAP1L2*, and *PRDM5*) show high mutation rate, copy number depletion, and/or decrease in expression in breast cancer (Supplemental Table S3). Based on the analysis of depleted gRNAs, we identified ERGs associated with cell cycle and cell senescence (Supplemental Fig. S14F), and these ERGs showed higher rate of copy number amplification or up-regulation in breast cancer (i.e., *ARID4B* and *EZH2*). Similarly, when the analysis was performed on A549 cells (e.g., comparison D0 vs. D14), we revealed 69 ERGs associated with enriched gRNAs (Supplemental Fig. S15), among which several genes showed high mutation, copy number alteration, or decrease in expression rates in lung cancer.

Finally, by overlapping the genes associated with enriched gRNAs or depleted gRNAs in A549 and MCF10A, we identified ERGs consistently implicated in proliferation in both cell types. Nine genes (*TET1*, *KDM1A*, *SMARCE1*, *IDH2*, *CBX3*, *BMI1*, *NAP1L1*, *ARID1B*, and *HDAC2*) were associated with enriched gRNAs in both cancer types (Fig. 6B). Three genes (*TET1*, *ARID1B*, and *BAZ1A*) are highly mutated in breast and lung cancer types (Supplemental Table S3), whereas *TET1*, *IDH2*, and *ARID1B* were also among the top genes altered in several cancer types (Supplemental Table S3), further corroborating their putative role as cancer drivers. A higher number of genes were identified when depleted gRNAs were overlapped between the two cell lines (Fig. 6C,D) several of which (i.e., *BOP1*, *RSF1*, *ACTL6A*, *ASH2L*, and *ATR*) showed high copy number amplification or increase in expression in breast and lung cancer, suggesting that those genes might play essential roles in cancer proliferation and viability.

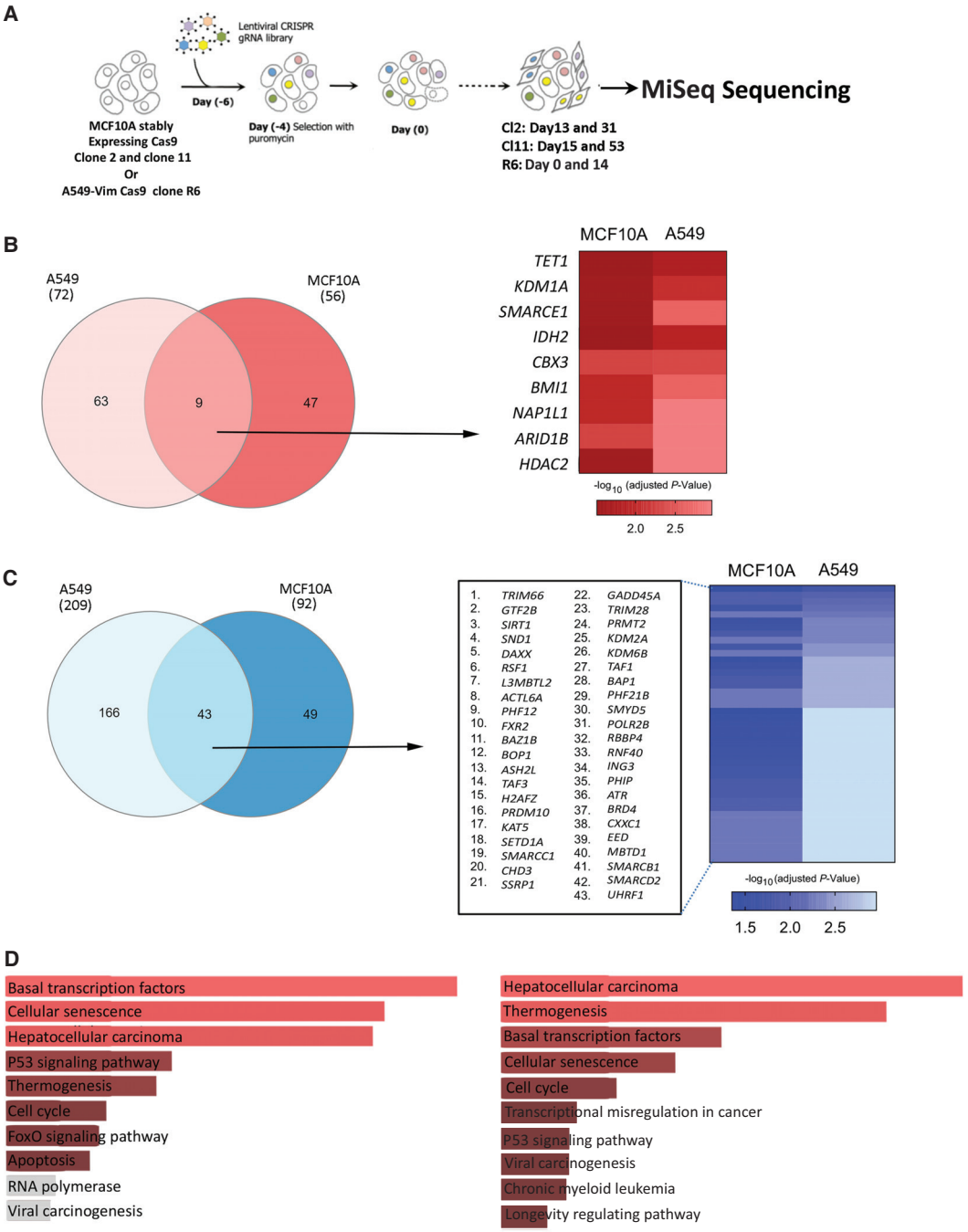


Figure 6. CRISPR-Cas9 screen to perform orthogonal assessment of the driver potential of ERGs in cancer cell proliferation. (*A*) The screening strategy used to identify regulators of cell proliferation among ERGs in both A549 and MCF10A cell lines/clones. (*B*,*C*, *left*) Venn diagrams showing the genes associated with significantly enriched (*B*) or depleted (*C*) gRNAs in the screens performed on A549 and MCF10A cells using edgeR analysis in CRISPRAnalyzeR. (*B*,*C*, *right*) Heatmaps showing the adjusted *P*-values of the commonly enriched (*B*) or depleted (*C*) gRNAs in both cell lines. Data are presented as $-\log_{10}$ (adjusted *P*-value). (*D*) KEGG pathway analysis performed on genes associated with commonly depleted gRNAs (*left*) and with commonly depleted and enriched gRNAs (*right*) in both cell lines. All pathways in red show $P < 0.05$.

Discussion

In the present study, we performed a pan-cancer analysis of (epi) genomic and transcriptomic alterations in a comprehensive panel of ERGs using available cancer genome data sets and a range of novel and powerful bioinformatics tools, revealing candidate epi-drivers across cancer types. We cataloged recurrent pan-cancer mutations or CNAs in specific ERGs or classes of ERGs. Application of driver prediction algorithms and orthogonal CRISPR-Cas9 in vitro screens revealed the ERGs with a potential driver role conferring

on cancer cells the traits associated with the hallmarks of cancer. This is the largest and most comprehensive analysis thus far of the cancer-associated disruption of ERGs and the first experimental effort to identify epidrivers in oncogenic processes through an ERG-wide screen.

Our finding that the predominant genetic alterations in ERGs across tumor samples of most cancer types are CNAs, notably amplifications, reveals that in addition to recurrent mutations, both amplifications and deletions in ERGs may play more important roles than previously anticipated. These results extend the previous studies on a limited set of cancer types showing that mutations in ERGs are recurrent (Gonzalez-Perez et al. 2013; Plass et al. 2013; Timp and Feinberg 2013; Vogelstein et al. 2013) and that amplified regions are enriched for genes involved in epigenetic regulation (Zack et al. 2013). The finding that some cancer types (such as OV and SARC) show frequent deep CNAs with virtually no SNAs argues that the roles of epidrivers in those cancer types may be driven primarily by CNAs, with a relatively minor role of mutations. Our results that different ERG classes showed similar patterns of genetic alterations, with the exception of DNA methylation writers, which showed a markedly high ratio of amplifications to deletions, suggest that amplification of this class may be the principal mechanism of their genetic deregulation across cancer types. These different patterns of alterations may reflect distinct mechanisms by which these CNAs are generated and/or positive/negative selection during tumor development.

The importance of CNAs highlighted herein further suggests that any driver prediction model, particularly for ERGs, would need to account for genetic amplifications and deletions. The inclusion of other omics phenotypes, such as RNA expression, is also important particularly given the high level of RNA expression aberrations observed in ERGs across many cancer types. Recent evidence questions the conventional interpretation of hotspot mutations as being evidence of positive selection and driver events (Hess et al. 2019), further reinforcing the need to integrate multiple omics in driver prediction models. Accordingly, we proposed the Multi-Omics and Pan-Cancer Driver prediction tools, which account for the SNAs, CNAs, and RNA expression aberrations and complement (rather than overlap with) the ConsensusDriver approach, which seems to be heavily weighted by SNA frequencies, at least for ERG driver prediction based on our results. Whereas our Multi-Omics Driver score highlighted the epidrivers within each class of malignancy, our Pan-Cancer Driver score reflects the recurrence of driver potential across cancer types.

We observed both cancer type–specific and cancer-wide genetic deregulation of ERGs. A subset of ERGs was frequently genetically altered across many malignancies (SNAs in the *KMT2A/B/C/D* family members and *ARID1A*, and deep CNAs in *BOP1* and *ATAD2* were each seen in several cancer types), consistent with the notion that disruption of some ERGs represents a shared driver mechanism operating across multiple cancer types. Little overlap was observed between the ERGs showing a high frequency of SNAs and those with a high frequency of CNAs (except for a few ERGs) (Figs. 2E–H, 4C). Similarly, in expression analysis, we observed both cancer type–specific and cancer-wide deregulation of ERGs. Whereas ERG expression correlated highly with CNAs, it correlated negatively with some SNAs and positively with others, a finding consistent with recent evidence indicating that interaction between driver mutations and transcription may be context dependent (Ding et al. 2018). Furthermore, some discrepancy between mutation and expression alterations in ERGs may be explained by the impact of nonmutational mechanisms on gene expression. This is supported by our observation that some cancer types, such as GBM, have a low burden of SNAs (Fig. 2E) and CNAs (Fig. 2F) in ERGs, in line with a low within-tumor variation in ERG expression (Fig. 3A), but a high number of ERGs with deregulated expression in tumor samples relative to adjacent normal tissue (Fig. 3E–G) and vice versa (e.g., STAD). Our analysis of DNA methylation and RNA-seq levels revealed that tumor-specific differentially methylated CpGs in promoter regions were negatively correlated with the expression of their corresponding genes, underscoring the notion that epigenetic inactivation could be an additional mechanism for aberrant expression of ERGs.

ERGs, including top predicted driver genes, were commonly enriched in four of ten cancer hallmarks: genome instability and mutation, evading growth suppressors, sustaining proliferative signaling, and enabling replicative immortality (Fig. 4F). The latter two hallmarks overlapped, respectively, with cell cycle (accelerator of proliferation) and cellular senescence (defined as irreversible cell cycle arrest, hence, a decelerator of proliferation), which were found to be among the top pathways deregulated in the genes associated with enriched gRNAs (Supplemental Figs. 14D and 15D). Although it has been proposed that the hallmarks of cancer are acquired through distinct mechanisms in different cancer types (Hanahan and Weinberg 2011), our results suggest that many of these functional capabilities may be acquired through the shared mechanism involving the disruption of ERGs.

Our orthogonal in vitro CRISPR-Cas9 screen also identified a set of specific ERGs affecting markers of tumorigenesis such as cell proliferative potential and EMT. Our results that five epidriver candidates (*SRCAP*, *EP400*, *ARID1B*, *MBD5*, and *KMT2A*) among the top 15 ERGs enriched in EMT fraction (Fig. 5H; Supplemental Tables S6, S9) were found among the most mutated ERGs in clinical samples across cancer types (Fig. 2G) support the driver role of EMT-specific ERG tumorigenesis. In addition, an analysis of the interaction networks of the top ERG hits associated with EMT (*KAT2B*, *EP400, SRCAP*, *ARID1B*, and *SUV39H1*) uncovered several directional dependencies involving genes known to play a role in EMT and multiprotein complexes regulating chromatin structure and function, connecting previously uncharacterized complexes/pathways to the EMT process.

This study contributes to a greater understanding of the deregulation of ERGs and their functional impact in cancer. This insight should prove instrumental in the clinical application of ERGs, especially considering a growing interest in developing epigenetics-based prognostic and therapeutic strategies. Developing "epigenetic drugs" capable of modulating specific ERGs (epidrivers) can circumvent the high toxicity and off-target effects of broad epigenome reprogrammers (DNMT inhibitors, histone deacetylase inhibitors) and offer a potent tool for precision medicine (Ahuja et al. 2016; Brien et al. 2016; Jones et al. 2016). Therefore, the results of our study may provide the basis for translational approaches aimed at developing epigenetics-based early detection and personalized cancer treatment and prevention.

Methods

Generating the compendium of epigenetic regulator genes

A compendium of genome-wide ERGs was generated by integrating the different available gene databases (The Human Gene Database—GeneCards, https://www.genecards.org/; the NCBI Eukaryotic Genome Annotation Pipeline https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/; Cytoscape version

3.6.1) and the most relevant publications (Gonzalez-Perez et al. 2013; Plass et al. 2013; Timp and Feinberg 2013; Vogelstein et al. 2013; Yang et al. 2015; Xu et al. 2017). This resulted in a comprehensive list of 426 genes, which we categorized into 12 groups based on their functional features: (1) Histone methylation editor (HM_e) = histone demethylases (HDMs); (2) histone methylation writer (HM_w) = histone methyltransferases (HMTs); (3) histone methylation reader (HM_r); (4) DNA methylation writer (DM_w); (5) DNA methylation editor (DM_e); (6) DNA methylation reader (DM_r); (7) histone acetylation editor (HA_e) = histone deacetylases (HDACs); (8) histone acetylation writer (HA_w) = histone acetyltransferases (HATs); (9) histone acetylation reader (HA_r); (10) chromatin remodeling complex (ChRC); (11) helicases; and (12) other chromatin modifiers = the remaining ERGs included in the study (Fig. 1B). To classify ERGs based on their potential function as a tumor suppressor or an oncogene, we used the TSGene (https://bioinfo.uth.edu/TSGene/) (Zhao et al. 2013) and OncoKB (https://oncokb.org/) (Chakravarty et al. 2017) databases, respectively.

Data resource

We downloaded the data sets using the publicly available TCGA (provisional) database from cBioPortal (https://www.cbioportal.org/datasets), which consists of data sets with genetic alterations, including single-nucleotide alterations (SNAs) and copy number alterations (CNAs), and gene expression (expression median and *Z*-scores) of 426 ERGs. For reproducibility, analyses were repeated using the TCGA expression and genetic data downloaded on May 16, 2019, and March 26, 2019, respectively.

We used TCGA abbreviations for 33 cancer types as follows: (LAML) acute myeloid leukemia; (ACC) adrenocortical carcinoma; (BLCA) bladder urothelial carcinoma; (LGG) brain lower grade glioma; (BRCA) breast invasive carcinoma; (CESC) cervical squamous cell carcinoma and endocervical adenocarcinoma; (CHOL) cholangiocarcinoma; (COAD) colon adenocarcinoma; (ESCA) esophageal carcinoma; (GBM) glioblastoma multiforme; (HNSC) head and neck squamous cell carcinoma; (KICH) kidney chromophobe; (KIRC) kidney renal clear cell carcinoma; (KIRP) kidney renal papillary cell carcinoma; (LIHC) liver hepatocellular carcinoma; (LUAD) lung adenocarcinoma; (LUSC) lung squamous cell carcinoma; (DLBC) lymphoid neoplasm diffuse large B cell lymphoma; (MESO) mesothelioma; (OV) ovarian serous cystadenocarcinoma; (PAAD) pancreatic adenocarcinoma; (PCPG) pheochromocytoma and paraganglioma; (PRAD) prostate adenocarcinoma; (READ) rectum adenocarcinoma; (SARC) sarcoma; (SKCM) skin cutaneous melanoma; (STAD) stomach adenocarcinoma; (TGCT) testicular germ cell tumors; (THYM) thymoma; (THCA) thyroid carcinoma; (UCS) uterine carcinosarcoma; (UCEC) uterine corpus endometrial carcinoma; (UVM) uveal melanoma.

Pan-cancer analysis of genetic alterations in ERGs

The proportions of each of the CNAs and SNAs detected in ERGs were calculated among tumor samples within each of the 33 cancer types. The raw SNA data set contained somatic, nonsynonymous mutations, which were transformed into data with mutation status indicating the number of SNAs for a gene in each sample. The raw CNA data set was used to characterize CNAs by genomic position and amplitude as follows: CNA = 0 indicates diploid with no alteration; amplification = 1 indicates a shallow gain (a few additional copies, often broad); deep amplification = 2 indicates a high-level amplification (more copies, often focal); shallow deletion = −1 indicates a shallow loss, possibly a heterozygous deletion; deep deletion= −2 indicates a deep loss, possibly a homozygous deletion (https://www.cbioportal.org/). For more robust analyses, we regrouped CNAs into three groups: (1) no alteration, (2) deep amplification (amp), and (3) deep deletion (del). We then pooled together the resulting merged data sets of SNAs and CNAs matched to the same sample ID. The same criteria were used for calculating genetic deregulation in all genes and for ERG functional classes. Circos plots were generated as described previously (Krzywinski et al. 2009; Gu et al. 2014).

Multi-omics analysis of genomic and transcriptomic aberrations in ERGs

To effectively visualize multidimensional data sets of the deregulation of ERGs across different cancer types, we integrated SNAs, CNAs, and RNA expression alterations. For each ERG, the proportions of CNAs versus SNAs were calculated among tumor samples within each of 33 cancer types. The analysis included only genetically deregulated genes that show SNAs and CNAs (amplification = 1, 2, and deletion = −1, −2) in at least 1% and 10% of the samples in any cancer, respectively. ERGs were then classified based on their mutation profiles such that those harboring mostly SNAs, with CNAs not passing the threshold of 10%, were considered as only mutated; whereas genes with CNAs in at least 10% of the samples were considered as amplified or deleted. Finally, pooled results of SNAs and CNAs were integrated with RNA expression data, expressed as *Z*-score values for each corresponding ERG in each cancer type. The percentage of samples with significant *Z*-scores ($Z > 2$ or $Z < -2$) was reported for each ERG in each cancer type.

Differential expression analysis of ERGs in tumor samples and adjacent normal tissues

We downloaded HTSeq count files from the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov) for each cancer type and divided adjacent normal samples and tumor samples based on the ID annotation of TCGA samples. Of the total of 33 cancer types, we focused on 18 cancer types that had available ID annotation for adjacent normal samples. We used DESeq2 analysis to calculate expression changes comparing tumor samples with adjacent normal tissues. Genes with coverage of fewer than 10 reads were excluded. Only ERGs with absolute values of log FC > 1 and FDR < 0.05 were considered to be significantly differentially expressed.

Co-occurrence and mutual exclusivity analysis

Co-occurrence and mutual exclusivity analysis was performed in each cancer type separately and then meta-analyzed across cancer types. The data set was limited to the samples that had information about both SNAs and CNAs from cBioPortal (nonsynonymous mutations, fusions, deep amplifications, and deep deletions). For each gene pair and cancer type, we calculated an odds ratio (OR) quantifying how strongly the presence or absence of SNAs and/or CNAs in the first gene was associated with the presence or absence of the alterations in the second gene. The *P*-values were derived from the ORs using the Fisher's exact test and were further adjusted for multiple testing using the Benjamini–Hochberg FDR correction. The Haldane-Anscombe correction was used to avoid a division-by-zero error. The significant ORs (FDR < 0.05) were averaged across cancer types for each gene pair. ORs greater or less than 1 indicate tendencies toward co-occurrence and mutual exclusivity, respectively. Specifically, within each co-occurrent gene pair, the proportion of samples with both genes mutated needed to represent at least 5%–10% of the samples per cancer type.

ERG driver prediction models

The characterization of potential driver roles for ERGs was based on ConsensusDriver, a novel approach that provides a systematic way to integrate the strengths of various driver prediction algorithms (Bertrand et al. 2018). ConsensusDriver scores for ERGs overlapping with the ConsensusDriver genes (Bailey et al. 2018) were shown as a heatmap. The Pan-Cancer Driver score was generated for each of the 426 ERGs using a ranking method that accounts for SNAs, CNAs, and RNA expression aberrations within each cancer type and across multiple cancer types (the script provided on GitHub: https://github.com/IARCbioinfo/EPIDRIVE R2020 and as Supplemental Code). First, in the SNA and CNA data, we focused on ERGs that had a SNA or CNA (sharing the same direction) in at least 5% of samples in any cancer type. Then, for each gene, we counted the number of cancer types in which that gene had a SNA or CNA; the number of cancer types was used as a primary ranking method, and the percentage of samples showing alterations was used as the secondary ranking for genes having identical primary ranks. The gene with the lowest rank was given a score of 1, the second rank a score of 2, and so on. Genes ($n = x$) with equal ranks "y" (based on both the primary and secondary methods) were all assigned the same ranking score "y"; a subsequent gene with rank "y + 1" was then assigned a ranking score of "y + x." Two rankings were obtained, one for SNA and, independently, another for CNA. For expression data, we used both the *Z*-score and $\log_{10}$FC data and similarly calculated the ranking by counting for each gene the number of cancer types in which that gene had a $|Z\text{-score}| > 2$ or $|\log_{10}\text{FC}| > 2$ (and then using a secondary ranking based on the exact proportion of samples with $|Z\text{-score}| > 2$ or on the exact values of $|\log_{10}\text{FC}|$). Two rankings were obtained, one for *Z*-score data and one for $\log_{10}$FC data. The resulting four rankings were combined into one, such that SNA and CNA rankings each had a weight of 1, whereas *Z*-score and $\log_{10}$FC each had a weight of 0.5 (because both *Z*-score and $\log_{10}$FC reflect expression aberration).

Hallmarks of cancer enrichment of ERGs

To investigate whether deregulated ERGs are enriched in pathways affecting the 10 hallmarks of cancer (Hanahan and Weinberg 2011), an analysis was done separately for all ERGs or for genetically altered ERGs (by SNAs and independently by CNAs). For each cancer type, we included only genes with CNA > 1% or SNA > 1% across samples within a given cancer type. Then, for each cancer type, we calculated the enrichment of its genetically deregulated genes in every hallmark using the Fisher's exact test and further adjusted for multiple testing using the Bonferroni correction.

Generation of cell clones stably expressing Cas9

A549 Vim RFP cells (ATCC CCL-185EMT) and MCF10A cells (ATCC CRL-10317) were cultured according to the recommendation by ATCC. To generate stably expressing Cas9 cell lines, A549 Vim RFP and MCF10A cells were transfected with lentiviral particles containing Cas9 nuclease (GeneCopoeia 217LPP-CP-LVC9NU-01-100-C) and Lenti-Cas9-2A-Blast plasmid (Addgene 73310), respectively, at a multiplicity of infection (MOI) of 5. Briefly, cells were cultured for 5 h in cell culture media supplemented with 8 µg/mL of polybrene. Spinoculation was then applied at 800*g* for 90 min at 37°C. At 48 h after transfection, 500 µg/mL G418 (cells transfected with 217LPP-CP-LVC9NU-01-100-C) or 10 µg/mL blasticidin (cells transfected with Lenti-2A-Cas9-blast) were used for positive selection of Cas9-transduced cells. For each cell type, we then generated single clones stably expressing Cas9 using cloning rings or serial dilutions followed by cultures of single cell. The expression of Cas9 protein in individual cell clones were determined by western blot analysis using Anti-CRISPR-Cas9 antibody (Abcam 7A9-3A3). A549 Vim RFP Cas9 clone R6 and MCF10A Cas9 clones 2 and 11 were used for the CRISPR-Cas9 library screenings.

Construction of the CRISPR-Cas9 sgRNA library and titration

We generated a CRISPR library comprising 1649 different gRNAs targeting 426 human ERGs. Each candidate gene was targeted by 1–4 sgRNA (Supplemental Table S7). Lentiviral plasmids containing sgRNAs were obtained in bacterial glycerol stock from a commercial source (Thermo Fisher Scientific). We pooled and amplified together 10 µL of glycerol stock of each plasmid gRNA to obtain a homogeneous representation of the library, followed by DNA extraction using Maxi prep (Qiagen). The library was packaged in human embryonic kidney HEK293T cells using a third-generation lentivirus expression system consisting of the mixture of 20 µg of the transfer vector consisting of the pool of sgRNAs lentiviral plasmid constructs (20 µg), 12.5 µg of the packaging plasmid I pMDLg/pRRE (Addgene 12251), 7.5 µg of the packaging plasmid II pRSV-Rev (Addgene 12253), and 7.5 µg of envelope plasmid VSV-G - pMD2.G (Addgene 12259) in Opti-MEM diluent. The library lentiviral particles were produced using the polyethylenimine method (Tebu-Bio). Two collected harvests were pooled together and concentrated using Lenti Concentrator (OriGene) according to the manufacturer's instructions. The resulting lentivirus CRISPR library was aliquoted and stored at −80°C. The virus titer and optimal transduction efficiency (considered to be 40%) of the lentivirus library were determined by colony formation assay in A549 Vim Cas9 cells.

Evaluation of sgRNAs representation in the generated library

To evaluate the relative representation of sgRNAs in the library, we performed deep sequencing using MiSeq (Illumina) (Supplemental Fig. S8). Briefly, we designed primers (forward, CGATACAAGGCT GTTAGAGAGATA; reverse, GTTGCTATTATGTCTACTATTCTTT CCC) to obtain a 430-bp amplicon of plasmid DNA containing the sgRNA sequences. We followed the suggestion of Illumina to have an amplicon length of >300 bp for the targeting sequencing, using the Nextera XT DNA Sample Preparation Kit (Illumina) according to the manufacturer's instructions. A single gRNA of the *AKAP1* gene was chosen from the candidate gene list to be evaluated as a positive control for library distribution. Targeted sequencing of the pooled library and a single gRNA of the *AKAP1* gene was performed using the MiSeq Reagent Kit v2 (500 cycles, Illumina) according to the manufacturer's instructions. The FastQC generated from MiSeq was analyzed in Galaxy using the BLASTN tool (2.5.0+ Package: blast 2.5.0). The sgRNAs were mapped against all sgRNA sequences present in the custom-made CRISPR library. The representation of genes in the pooled library was calculated relative to the abundance of gRNA for each gene (Supplemental Fig. S8).

CRISPR-Cas9 library screening for epidrivers of epithelial-to-mesenchymal transition

A549 Vim RFP Cas9 cells were transduced with the lentivirus CRISPR library at a MOI of 0.3. For a negative control, we used the nontargeting sgRNA LentiArray CRISPR Negative Control Lentivirus (Thermo Fisher Scientific). The baseline time point (day 0) was designated as 4 d after puromycin selection (the time point when untransduced A549 Vim RFP Cas9 cells were dead). The library was applied in two technical duplicates and in two independent experiments. During cell passaging, $\sim 2 \times 10^6$ cells were

maintained in culture to achieve on average 1000-fold coverage of all 1649 sgRNAs in the library. We isolated the EMT population at days 14, 21, and 28 using FACS sorting (S3e, Bio-Rad) for RFP (Vimentin)-positive cells (Vimentin positive [VIM$^+$]).

Validation of the isolated EMT population

RNA of control A549 Vim RFP Cas9 cells and the sorted VIM$^+$ cells (from all time points) were extracted using the AllPrep DNA/RNA Mini Kit (Qiagen). To validate the EMT transition in VIM$^+$ cells, mRNA expression levels of several EMT markers were analyzed by quantitative RT-PCR. Cadherin 1/cadherin 2 ratio and expression levels of SNAI1 (also known as SNAIL) and ZEB2 were determined and confirmed the EMT state of VIM$^+$ cells (Supplemental Fig. S16). Fluorescence microscopy was used to verify RFP Vimentin-positive (red) cells (Supplemental Fig. S17) in VIM$^+$ sorted cells compared to the parental cell lines. For additional validation of the isolated EMT population, intensity levels of the epithelial marker EPCAM (Fig. 5E; Supplemental Fig. S9) and the mesenchymal marker cadherin 2 (Fig. 5F) were determined using flow cytometry (S3e, Bio-Rad) using EPCAM-APC antibody (Miltenyi Biotec 130-111-000) and cadherin 2 antibody (R&D systems IC1388P).

CRISPR-Cas9 library screening for epidrivers of cell proliferation

A549 Vim Cas9 Clone R6 and MCF10A Cas9 Clone 2 and Clone 11 cells were transduced with the lentiviral CRISPR library of 1649 sgRNAs (MOI 0.3 and 0.1, respectively) and maintained in culture for several passages. Cells were collected at two different time points per cell line throughout their culture. The A549 cells were collected at day 0 (the day following the end of puromycin selection, early time point) and day 14 (late time point). Because the proliferation rate of MCF10A cells is lower than A549, we collected MCF10A cells at later passages. Two independent clones expressing Cas9 were used for MCF10A: Clone 2 and Clone 11 that were collected at early time points (days 13 and 15, respectively) and later time points (day 31 and 53, respectively) following transduction and puromycin selection. To achieve 1000-fold coverage of all sgRNAs, $\sim 2 \times 10^6$ cells were kept in culture at each passage, and 2×10^6 were collected at each time point.

Identification of enriched and depleted sgRNAs and their associated candidate epidrivers

To identify the enriched and depleted sgRNAs in the CRISPR-Cas9 screens, genomic DNA was isolated from 2×10^6 cells of each of the cell populations analyzed (VIM$^+$ cells and transduced parental cell lines at different time points) using AllPrep DNA/RNA Mini Kit. DNA was subsequently subjected to PCR to amplify the same region as that used for the validation of library representation (see above), containing sgRNA sequences using NEBNext High-Fidelity 2X PCR Master Mix (Illumina). PCR products were used for library preparation using the Nextera XT DNA Sample Preparation Kit or Nextera DNA Flex library Prep (Illumina) according to the manufacturer's instructions and sequenced on an Illumina MiSeq platform. The FastQC data generated by MiSeq were first analyzed in Galaxy using the BLASTN tool (2.5.0+ Package: blast 2.5.0). To obtain read counts, we performed mapping of sgRNA against all sgRNA sequences present in the custom-made CRISPR library. To identify the enriched and depleted sgRNAs in EMT epidriver screening, we performed paired analysis comparing the sorted VIM$^+$ populations at different time points (days 14, 21, and 28) with cells collected at day 0. The differential sgRNA abundance of the read counts was analyzed using CRISPRAnalyzeR software (DKFZ, Version: 1.50) (Winter et al. 2016). The list of enriched and depleted sgRNAs for each time point was defined by hit candidate overlapping the list of hit candidates identified by DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010; McCarthy et al. 2012) analyses (two independent analysis methods) with statistically significant changes of 0.001 and 0.01, respectively. The EMT candidate genes were identified as the genes commonly associated with enriched or depleted sgRNAs across different time points (days 14, 21, and 28) (Fig. 5G). Significantly enriched and depleted sgRNAs in the screening for epidrivers of cell proliferation were analyzed by paired analyses comparing sgRNAs detected at late time points with sgRNAs at early time point for each cell type using edgeR and/or DESeq2 statistical analyses methods in CRISPRAnalyzeR. Statistical significance was set at $P<0.001$ and $P<0.01$ for DESeq2 and edgeR, respectively.

Generation of single targeted knockout clones for EMT-identified epidrivers candidates

A549 Vim RFP Cas9 clone R6 was transfected with a pool of four sgRNAs targeting the genes of interests: *EP400*, *MBD5*, *ARID1B*, or *KAT2B*. Sequences of sgRNAs used are available in Supplemental Table S7. Transfection was performed using Xfect Transfection Reagent (Takara Bio) and 2 µg of pooled sgRNAs according to the manufacturer's instructions. Cells were subjected to puromycin (1 µg/mL) selection 36 h after transfection. After antibiotic selection, single clones were generated from the heterogeneous population of transfected resistant cells by amplification of single cell sorted by flow cytometry (FACSAria). To validate the alterations in the generated single clones, we designed PCR primer amplifying the regions surrounding the Cas9 cutting sites for each of the targeted genes. Targeted regions were amplified from genomic DNA extracted from the generated single clones, and PCR products were sequenced (Sanger sequencing) (Supplemental Fig. S18; Supplemental Table S8). Several alignment tools were used to analyse the sequencing data (CRISPR-ID: http://crispid.gbiomed.kuleuven.be; DSDecodeM: http://skl.scau.edu.cn/dsdecode; and https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Quantitative RT-PCR

Total RNA extraction was performed by using AllPrep DNA/RNA Mini Kit (Qiagen) according to the manufacturer's protocol using primers shown in Supplemental Table S8.

Scratch wound healing assay

A549 Vim Cas9 R6 parental cell clone and all generated A549 single targeted knockout clones were analyzed by scratch wound healing assay using a standard protocol. Briefly, cells were plated in 12-well plates (2×10^5 cells/well); after reaching 90%–100% confluence (~24 h after plating), two scratches per well were performed (in the form of a cross) using a 200-µL tip. Experiments were performed in duplicates. To measure cell migration and wound healing capacity, four different pictures were taken per well using a Zeiss TELAVAL 31 microscope and a Nikon D40 camera, both at 0-h (time of scratch) and 24-h time points. Closure of the wound by cell migration was calculated by comparing the scratched areas at both time points, using the ImageJ software (version 1.52b).

Transwell migration assay

The migratory properties of A549 Vim Cas9 R6 parental cell clone and all generated A549 single targeted knockout clones were analyzed by transwell migration assay. Experiments were performed in duplicates for each clone using cell culture polycarbonate 8-µm

inserts (Millicell) in 24-well plated (corning). Briefly, 1×10^4 cells in 300 µL of serum free F-12K medium were added to the upper part of cell inserts. To stimulate the migration, 500 µL of F-12K medium containing FBS were added to the lower part of the inserts. Cells were incubated for 13 h at 37°C and 5% CO_2. Thirteen hours is an insufficient time for A549 Vim Cas9 R6 parental cells to migrate. After incubation, the medium was carefully aspirated from the inside of the insert. The interior of the inserts was then gently swabbed with cotton-tipped swabs to remove nonmigratory cells. To stain migratory cells, the insert was transferred to a clean well containing 400 µL of crystal violet Cell Stain Solution and incubated for 20 min. After 2× washes with PBS, inserts were air dried and images taken by Zeiss TELAVAL 31 microscope and Nikon D40 camera.

Assessing disruption of EMT-specific driver candidates in clinical samples and pathway enrichment

To assess the disruption of EMT-specific driver candidates identified in the CRISPR-Cas9 screen in clinical samples, we analyzed the TCGA data using the Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/). We divided samples into nonmetastatic (M0) and metastatic (M1) subsets based on the American Joint Committee on Cancer metastasis stage code. For more robust analysis, we further selected samples based on tumor stage using the American Joint Committee on Cancer neoplasm disease stage code. Only stage I and IV were considered to be M0 and M1, respectively. Overall, we collected 873 M0 and 371 M1 cases across 21 different cancer types available in TCGA (Supplemental Table S9). Based on these data sets, we calculated the mean percentage of mutations of each of the EMT-specific ERGs in the M1 versus M0 subsets.

We performed the pathway enrichment analyses using bioinformatics mapping tools of different databases, including the NCI-Nature 2016, Panther 2016, KEGG 2016, and Reactome 2016 databases. We used Enrichr (Kuleshov et al. 2016) score of each database to define the pathway enrichment of EMT-specific epidrivers. The network of EMT-specific ERGs was constructed with the GeneMANIA package (https://genemania.org/) (Warde-Farley et al. 2010).

Data access

All raw sequencing data generated in this study have been deposited in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession number: PRJNA655831. The scripts used to generate Driver Scores were provided on GitHub (https://github.com/IARCbioinfo/EPIDRIVER2020) and as Supplemental Code.

Competing interest statement

The authors declare no competing interests.

Acknowledgments

We thank Dr. Maria Ouzounova for her help with flow cytometry and Mr. Cyrille Cuenin for his excellent technical support with lentiviral library preparation and high-throughput sequencing. We also thank Drs. Hector Hernandez-Vargas, Davide Degli-Esposti, Srikant Ambatipudi, Fazlur R. Talukdar, and Nora Fernandez Jimenez for their valuable input to initial stages of the project. We thank Dr. Claire Renard for her help with depositing the raw sequencing data on the NCBI database. We thank Drs. Michael Korenjak, Magali Olivier, Vlatka Zoldoš, Aleksandar Vojta, Rabih Murr, Lynnette Fernandez-Cuesta, and Nino Sinčić for critical reading of the manuscript; and Karen Müller for editing the manuscript. This work was partially supported by the grant from the Institut National du Cancer (INCa, France), La direction générale de l'offre de soins (DGOS), and Institut National de la Santé et de la Recherche Médicale (INSERM) (SIRIC LYriCAN, INCa-DGOS-Inserm_12563). Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy, or views of the International Agency for Research on Cancer/World Health Organization.

Author contributions: Conceptualization, Z.H.; methodology, Z.H., A.G., V.C., A.N., A.H., and R.K.; investigation, Z.H., A.G., V.C., A.N., M.G.S.A., R.K., and A.H.; writing—original draft, Z.H. and A.G.; writing—review and editing, Z.H., A.G., A.H., and R.K.; funding acquisition, Z.H.; resources, Z.H., A.G. V.C., A.N., and A.H.; supervision, Z.H. and A.G. All authors reviewed the manuscript.

References

Ahuja N, Sharma AR, Baylin SB. 2016. Epigenetic therapeutics: a new weapon in the war against cancer. *Annu Rev Med* **67:** 73–89. doi:10.1146/annurev-med-111314-035900

Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, Colaprico A, Wendl MC, Kim J, Reardon B, et al. 2018. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173:** 371–385.e18. doi:10.1016/j.cell.2018.02.060

Bell O, Tiwari VK, Thomä NH, Schübeler D. 2011. Determinants and dynamics of genome accessibility. *Nat Rev Genet* **12:** 554–564. doi:10.1038/nrg3017

Bertrand D, Drissler S, Chia BK, Koh JY, Li C, Suphavilai C, Tan IB, Nagarajan N. 2018. ConsensusDriver improves upon individual algorithms for predicting driver alterations in different cancer types and individual patients. *Cancer Res* **78:** 290–301. doi:10.1158/0008-5472.CAN-17-1345

Brien GL, Valerio DG, Armstrong SA. 2016. Exploiting the epigenome to control cancer-promoting gene-expression programs. *Cancer Cell* **29:** 464–476. doi:10.1016/j.ccell.2016.03.007

Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. 2017. OncoKB: a precision oncology knowledge base. *JCO Precis Oncol* **1:** 1–16. doi:10.1200/PO.17.00011

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang KL, Tokheim C, et al. 2018. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173:** 305–320.e10. doi:10.1016/j.cell.2018.03.033

Gonzalez-Perez A, Jene-Sanz A, Lopez-Bigas N. 2013. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. *Genome Biol* **14:** r106. doi:10.1186/gb-2013-14-9-r106

Gu Z, Gu L, Eils R, Schlesner M, Brors B. 2014. Circlize implements and enhances circular visualization in R. *Bioinformatics* **30:** 2811–2812. doi:10.1093/bioinformatics/btu393

Hanahan D, Weinberg AR. 2011. Hallmarks of cancer: the next generation. *Cell* **144:** 646–674. doi:10.1016/j.cell.2011.02.013

Hess JM, Bernards A, Kim J, Miller M, Taylor-Weiner A, Haradhvala NJ, Lawrence MS, Getz G. 2019. Passenger hotspot mutations in cancer. *Cancer Cell* **36:** 288–301.e14. doi:10.1016/j.ccell.2019.08.002

Jones PA, Issa JP, Baylin S. 2016. Targeting the cancer epigenome for therapy. *Nat Rev Genet* **17:** 630–641. doi:10.1038/nrg.2016.93

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Res* **19:** 1639–1645. doi:10.1101/gr.092759.109

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A, et al. 2016. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44:** W90–W97. doi:10.1093/nar/gkw377

Lahouel K, Younes L, Danilova L, Giardiello FM, Hruban RH, Groopman J, Kinzler KW, Vogelstein B, Geman D, Tomasetti C. 2020. Revisiting the tumorigenesis timeline with a data-driven generative model. *Proc Natl Acad Sci* **117:** 857–864. doi:10.1073/pnas.1914589117

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15:** 550. doi:10.1186/s13059-014-0550-8

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res* **40:** 4288–4297. doi:10.1093/nar/gks042

Meyerson M, Gabriel S, Getz G. 2010. Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* **11:** 685–696. doi:10.1038/nrg2841

Murr R, Loizou JI, Yang YG, Cuenin C, Li H, Wang ZQ, Herceg Z. 2006. Histone acetylation by Trrap–Tip60 modulates loading of repair proteins and repair of DNA double-strand breaks. *Nat Cell Biol* **8:** 91–99. doi:10.1038/ncb1343

Ng PK, Li J, Jeong KJ, Shao S, Chen H, Tsang YH, Sengupta S, Wang Z, Bhavana VH, Tran R, et al. 2018. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33:** 450–462.e10. doi:10.1016/j.ccell.2018.01.021

Parmigiani G, Boca S, Lin J, Kinzler KW, Velculescu V, Vogelstein B. 2009. Design and analysis issues in genome-wide somatic mutation studies of cancer. *Genomics* **93:** 17–21. doi:10.1016/j.ygeno.2008.07.005

Plass C, Pfister SM, Lindroth AM, Bogatyrova O, Claus R, Lichter P. 2013. Mutations in regulators of the epigenome and their connections to global chromatin patterns in cancer. *Nat Rev Genet* **14:** 765–780. doi:10.1038/nrg3554

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26:** 139–140. doi:10.1093/bioinformatics/btp616

Sawan C, Vaissière T, Murr R, Herceg Z. 2008. Epigenetic drivers and genetic passengers on the road to cancer. *Mutat Res* **642:** 1–13. doi:10.1016/j.mrfmmm.2008.03.002

Shen H, Laird PW. 2013. Interplay between the cancer genome and epigenome. *Cell* **153:** 38–55. doi:10.1016/j.cell.2013.03.008

Tam WL, Weinberg RA. 2013. The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nat Med* **19:** 1438–1449. doi:10.1038/nm.3336

Timp W, Feinberg AP. 2013. Cancer as a dysregulated epigenome allowing cellular growth advantage at the expense of the host. *Nat Rev Cancer* **13:** 497–510. doi:10.1038/nrc3486

Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. 2013. Cancer genome landscapes. *Science* **339:** 1546–1558. doi:10.1126/science.1235122

Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, et al. 2010. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res* **38:** W214–W220. doi:10.1093/nar/gkq537

Winter J, Breinig M, Heigwer F, Brügemann D, Leible S, Pelz O, Zhan T, Boutros M. 2016. Carpools: an R package for exploratory data analysis and documentation of pooled CRISPR/Cas9 screens. *Bioinformatics* **32:** 632–634. doi:10.1093/bioinformatics/btv617

Xu Y, Zhang S, Lin S, Guo Y, Deng W, Zhang Y, Xue Y. 2017. WERAM: a database of writers, erasers and readers of histone acetylation and methylation in eukaryotes. *Nucleic Acids Res* **45:** D264–D270.

Yang Z, Jones A, Widschwendter M, Teschendorff AE. 2015. An integrative pan-cancer-wide analysis of epigenetic enzymes reveals universal patterns of epigenomic deregulation in cancer. *Genome Biol* **16:** 140. doi:10.1186/s13059-015-0699-9

Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B, Lawrence MS, Zhsng CZ, Wala J, Mermel CH, et al. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* **45:** 1134–1140. doi:10.1038/ng.2760

Zhao M, Sun J, Zhao Z. 2013. TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res* **41:** D970–D976. doi:10.1093/nar/gks937

Received July 7, 2020; accepted in revised form August 26, 2020.

Pan-cancer multi-omics analysis and orthogonal experimental assessment of epigenetic driver genes

Andrea Halaburkova, Vincent Cahais, Alexei Novoloaca, et al.

Genome Res. published online September 22, 2020
Access the most recent version at doi:10.1101/gr.268292.120

VI. Results

In this part, we present results yielded from the three-way modelling, encompassing DNA methylation markers of BW, gestational age and child sex in relation to CL risk. Part of this work has been published (A. Finalized papers) and some ongoing (B. Ongoing work). The ongoing work also comprises a hypothesis-free approach representing an agnostic methylome-wide analysis in relation to ALL risk, and which could be complementary (hence, not necessarily yielding overlapping results) to the three-way modelling. Due to its agnostic nature, the hypothesis-free method is bound by limitations in statistical power and inability to sufficiently correct for potential confounders, hence, requiring the inclusion of larger sample sizes and additional cohorts which are in the scope of future follow up studies originating from the thesis work.

A. Relevant publications

The first three papers hereafter describe the results for the association of each of the three closely related intrinsic factors (BW (69), gestational age (157) and child sex (in preparation) and DNA methylation. They represent major findings that provide new insights into epigenetic mechanisms of early-life factors. In each study, we meta-analyzed EWAS results from multiple international birth cohorts and found profound associations with DNA methylation at birth for each of these intrinsic factors. Specifically, DNA methylation in neonatal blood was associated after Bonferroni correction with birthweight at 914 CpG sites, with gestational age at 8899 CpGs sites and with child sex at 46,554 CpG sites on autosomes and 9,372 on the X chromosome. We observed a difference in BW ranging from -183 to 178 grams per 10% increase in methylation. For GA, the largest association represents 2.5% methylation change per additional gestational week. For child sex, the differences in methylation levels were generally small with a median difference of 0.5%. Methylation levels of the CpGs identified in newborns were further investigated in older ages, and we identified that a substantial proportion of signals fades off in childhood and adolescence for BW and gestational age, unlike the CpG markers of child sex, the majority of which persisted throughout childhood. Functional impact of methylation changes on gene expression was observed for 84 BW CpGs and 246 gestational age CpGs. Furthermore, gestational age methylation markers in cord blood were also found significant in lung and brain, highlighting that the cord blood findings capture the epigenomic plasticity of pre-natal development across tissues. Our access to biospecimens and data was made possible by groundwork undertaken to establish the collaborations and resources. These three projects are part of large consortia that require significant amount of data harmonization, analysis and scrutiny as well as routine follow-up with several cohorts worldwide. Manuscript writing often requires substantially more time in these consortium workflows than is the case for other types of studies we have.

In the fourth paper (158), we used the significant BW-associated CpGs identified in our previous study (69) and integrated four types of omic data (methylome, transcriptome, metabolome and a set of inflammatory proteins) measured in cord

blood samples from four birth-cohorts in order to investigate in further depth the molecular mechanisms that underlie changes in BW. Multi-omic analysis revealed interrelated epigenetic, transcriptional, metabolic and protein pathways converging on an important role for cholesterol biosynthesis in affecting BW. This finding was further reinforced by a significant association detected between measured cord blood cholesterol levels and BW. Further studies are required to determine the causality of these findings and their role on child health.

1) Küpers LK, Monnereau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, Page CM, Reese SE, Wilcox AJ, Czamara D, Starling AP, **Novoloaca A**, Lent S, Roy R, Hoyo C, Breton CV, Allard C, Just AC, Bakulski KM, Holloway JW, Everson TM, Xu CJ, Huang RC, van der Plaat DA, Wielscher M, Merid SK, Ullemar V, Rezwan FI, Lahti J, van Dongen J, Langie SAS, Richardson TG, Magnus MC, Nohr EA, Xu Z, Duijts L, Zhao S, Zhang W, Plusquin M, DeMeo DL, Solomon O, Heimovaara JH, Jima DD, Gao L, Bustamante M, Perron P, Wright RO, Hertz-Picciotto I, Zhang H, Karagas MR, Gehring U, Marsit CJ, Beilin LJ, Vonk JM, Jarvelin MR, Bergström A, Örtqvist AK, Ewart S, Villa PM, Moore SE, Willemsen G, Standaert ARL, Håberg SE, Sørensen TIA, Taylor JA, Räikkönen K, Yang IV, Kechris K, Nawrot TS, Silver MJ, Gong YY, Richiardi L, Kogevinas M, Litonjua AA, Eskenazi B, Huen K, Mbarek H, Maguire RL, Dwyer T, Vrijheid M, Bouchard L, Baccarelli AA, Croen LA, Karmaus W, Anderson D, de Vries M, Sebert S, Kere J, Karlsson R, Arshad SH, Hämäläinen E, Routledge MN, Boomsma DI, Feinberg AP, Newschaffer CJ, Govarts E, Moisse M, Fallin MD, Melén E, Prentice AM, Kajantie E, Almqvist C, Oken E, Dabelea D, Boezen HM, Melton PE, Wright RJ, Koppelman GH, Trevisi L, Hivert MF, Sunyer J, Munthe-Kaas MC, Murphy SK, Corpeleijn E, Wiemels J, Holland N, Herceg Z, Binder EB, Davey Smith G, Jaddoe VWV, Lie RT, Nystad W, London SJ, Lawlor DA, Relton CL, Snieder H, Felix JF. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 2019 Apr 23;10(1):1893.
2) Merid SK*, **Novoloaca A***, Sharp GC*, Küpers LK*, Kho AT*, Roy R, Gao L, Annesi-Maesano I, Jain P, Plusquin M, Kogevinas M, Allard C, Vehmeijer FO, Kazmi N, Salas LA, Rezwan FI, Zhang H, Sebert S, Czamara D, Rifas-Shiman SL, Melton PE, Lawlor DA, Pershagen G, Breton CV, Huen K, Baiz N, Gagliardi L, Nawrot TS, Corpeleijn E, Perron P, Duijts L, Nohr EA, Bustamante M, Ewart SL, Karmaus W, Zhao S, Page CM, Herceg Z, Jarvelin MR, Lahti J, Baccarelli AA, Anderson D, Kachroo P, Relton CL, Bergström A, Eskenazi B, Soomro MH, Vineis P, Snieder H, Bouchard L, Jaddoe VW, Sørensen TIA, Vrijheid M, Arshad SH, Holloway JW, Håberg SE, Magnus P, Dwyer T, Binder EB, DeMeo DL, Vonk JM, Newnham J, Tantisira KG, Kull I, Wiemels JL, Heude B, Sunyer J, Nystad W, Munthe-Kaas MC, Räikkönen K, Oken E, Huang RC, Weiss ST, Antó JM, Bousquet J, Kumar A, Söderhäll C, Almqvist C, Cardenas A, Gruzieva O, Xu CJ, Reese SE, Kere J, Brodin P, Solomon O, Wielscher M, Holland N, Ghantous A, Hivert MF, Felix JF, Koppelman GH, London SJ, Melén E. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. Genome Med. 2020 Mar 2;12(1):25.

*Co-first authorship.

3) Solomon O, Huen K, Kupers LK, Suderman M, Gruzieva O, Gao L, Bakulski KM, **Novoloaca A**, Allard C, Pappa I, Yousefi P, Llambrich M, Vives M, Jima DD, Kvist T, Baccarelli A, Sharp GC, Eskenazi B, Bergström A, Dou JF, Isaevska E, Corpeleijn E, Perron P, Jaddoe VWV, Nøhr E, Maitre L, Foraster M, Hoyo C, Lahti J, DeMeo DL, Kull I, Feinberg JI, Gagliardi L, Bouchard L, Tiemeier H, Santorelli G, Maguire RL, Czamara D, Litonjua A, Plusquin M, Lepeule J, Binder E, Dwyer T, Carracedo A, Raikkönen K, Kogevinas M, Nawrot TS, Munthe-Kaas MC, Herceg Z, Relton C, Melén E, Breton C, Fallin MD, Ghantous A, Snieder H, Hivert MF, Felix JF, Sørensen TIA, González JR, Bustamante M, Murphy SK, Oken E, London SJ, Holland N. Meta-analysis of epigenome-wide association studies in newborns and children show widespread sex differences in blood DNA methylation Child sex and DNA methylation. In preparation.
4) Alfano R, Chadeau-Hyam M, Ghantous A, Keski-Rahkonen P, Chatzi L, Perez AE, Herceg Z, Kogevinas M, de Kok TM, Nawrot TS, **Novoloaca A**, Patel CJ, Pizzi C, Robinot N, Rusconi F, Scalbert A, Sunyer J, Vermeulen R, Vrijheid M, Vineis P, Robinson O, Plusquin M. A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism. Metabolism. 2020 Sep;110:154292.

B. Ongoing Work

1. Three-way modelling

Exploring epigenetic mechanisms linking early-life factors to disease outcomes requires sufficient numbers of individuals that are both exposed and diseased. Hence, this can be challenging in a one-cohort setting, particularly when disease outcomes are rare, so international collaboration becomes crucial. We proposed a three-way modelling to address these challenges, and we test it on our hypothesis stipulating that DNA methylation deregulation *in utero* underlies the biological pathways linking early-life factors and CL. The approach consists of three consecutive steps as shown in the **Figure 7** and detailed below:

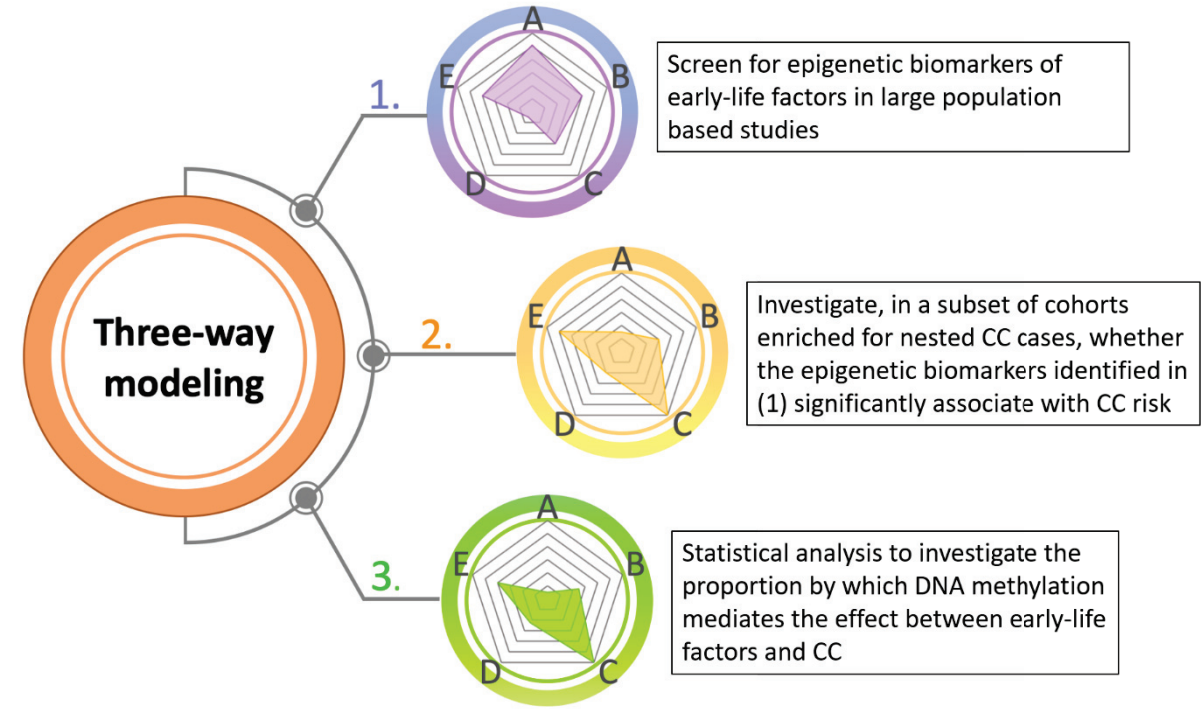


Figure 7. *Three-way modelling. The vertices in each radar chart correspond to: A – Number of tests; B – Sample size; C – Proportion of exposed samples; D – Proportion of diseased samples; E – Statistical power.*

First, CpG biomarkers of early-life factors are screened for in large population-based studies (abundant non-diseased subjects) using an epigenome-wide coverage (herein, the large sample sizes of PACE provide statistical power for high-coverage methylome profiling and adjustment for several confounders). We have applied this step on each of BW (69), gestational age (157) and child sex, which represent inter-related intrinsic factors in association to exposure and CC risk (see section IV). For the next steps, we will focus on the 8696 FDR (False Discovery Rate) significant methylation markers of BW and assess if they associate to CC risk and are affected by gestational age and child sex.

Second, the specifically significant CpGs obtained in the first step (reduced epigenetic dimension) are analyzed in a (smaller) subset of samples that are enriched for nested cancers using generalized linear models and adjusting for gestational age, child sex and maternal smoking. Herein, statistical power is maintained due to a targeted lower coverage profiling focusing on specific CpG markers. For this purpose, we used samples from the Norwegian Mother, Father and Child Cohort Study (MoBa) representing one of the two largest CC birth cohorts worldwide and encompassing prospectively collected DNA methylation data from 22 preB-ALL cases and 180 controls.

Third, statistical modelling such as mediation analysis (**Figure 8**) is used to investigate whether the identified CpG biomarkers mediate the effect of exposure

(BW) on cancer outcome (ALL risk). For this step, we applied two different mediation techniques that are subsequently described in more detail.

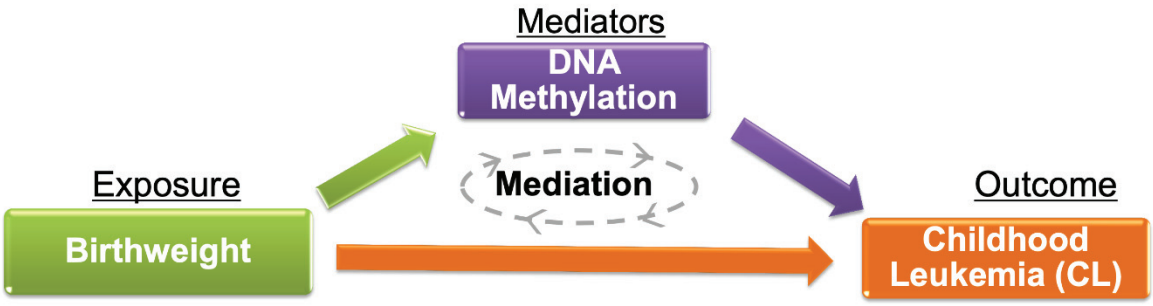
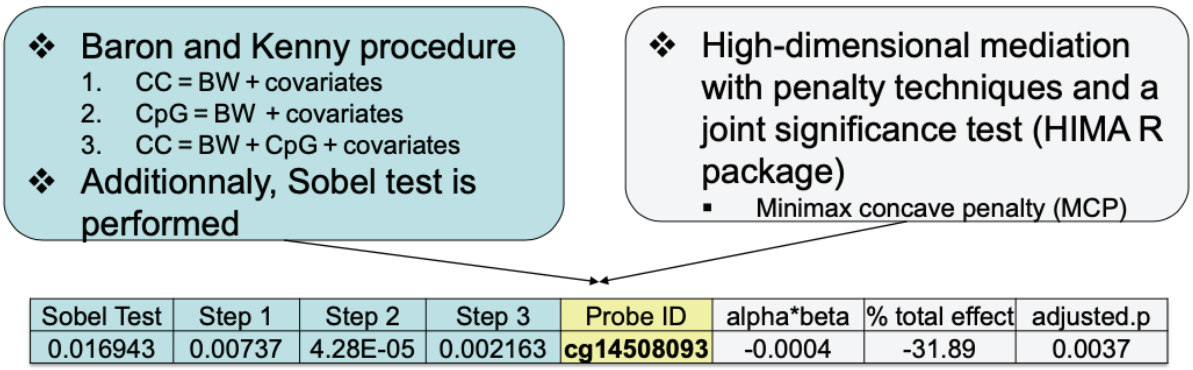


Figure 8. *Mediation analysis*

The first mediation approach (159) is studying a single mediator at a time. The well-known Baron and Kenny procedure is used herein and consists of three consecutive steps when we verify the establishment of (i) an effect that may be mediated between the exposure and the outcome, (ii) an effect between the exposure and the mediator and finally (iii) an effect between the mediator and the outcome. Additionally, we performed the Sobel test, which is more accurate than the Baron and Kenny procedure in case of large sample sizes (160). However, as the Sobel test is based on normal distributions, small samples sizes or distribution skewness can compromise the method's performance. For this reason, it is common practice to use both tests in mediation analyses.

The second approach (161) is a high-dimensional mediation analysis (HIMA) conducted with a penalty (162) and a joint significance test for mediation effect. The use of this second method is motivated by the fact that Baron and Kenny procedure was designed for one single mediator, and this approach tries to deal with multiple mediators which is the case in our study. Additionally, the HIMA algorithm was implemented in a ready-to-use R package.

To investigate the associations between the epigenetic markers of BW and CC status, we used a generalized linear model with adjustment for covariates: gestational age, gender, WBC composition and maternal smoking (known to have a strong effect on both the newborn's epigenome (118) and BW). Among the 8696 CpGs of BW, 414 were associated to CL (p.value < 0.05). Gene ontology enrichment of top most significant associations involves organ development pathways. Finally, both mediation techniques converged onto one CpG in a non-coding gene that mediates the association between BW and CL (adjusted p = 0.0037) and is not confounded by gestational age or child sex (**Figure 9**). Validating these findings requires replication of the analysis in other cohorts and experimental verification through functional assays.



Sobel Test	Step 1	Step 2	Step 3	Probe ID	alpha*beta	% total effect	adjusted.p
0.016943	0.00737	4.28E-05	0.002163	**cg14508093**	-0.0004	-31.89	0.0037

***Figure 9**. Common CpG between the two mediation techniques*

This work (in preparation) may add a new dimension to the value of EWAS findings by extending its current '2D modelling' (ie exposure- methylation or methylation-outcome) to another level of '3D analysis', hence, affording new opportunities for identifying robust epigenetic markers of exposure that function as precursors of disease endpoints, including rare ones.

2. DNA Methylation and ALL

Another complementary step was testing an agnostic approach to potentially identify differentially methylated regions between cases and controls in neonatal blood before the onset of ALL. This hypothesis-free investigation refers to the purple axis in the **Figure 6**. We used data from two studies, one from I4C (being MoBa) and one from CLIC (being California Childhood Leukemia Study (CCLS)) in a meet-in-the-middle framework that brings together the prospective and retrospective settings. We used neonatal bloodspots from CLIC collected retrospectively from archived samples and from which 450K data was generated. Neonatal methylome data from archived blood spots can capture a molecular snapshot of the early-life epigenome before cancer occurs, hence, making them potentially comparable to the prospective setting of I4C (which we attempt to demonstrate in this work as a proof-of-concept). To increase the statistical power of this agnostic methylome-wide approach, advanced bioinformatics and statistical modelling were implemented in order to reduce the dimension of the methylome into clusters of correlated and genomically proximal DNA methylation (CpG) sites, as per our previous work (154,155). Using the DMRcate statistical framework, we identified only a few (n = 4) differentially methylated regions, each spanning 3 to 16 CpG methylation sites, in neonatal blood before the onset of the leukemia in cases versus controls. Biological replication was successfully performed in a third study based on archived neonatal blood spots from the Melbourne study, Australia. The small number of significant results using this approach is expected due to the limited statistical power of this design. However, this pilot work suggests that archived neonatal blood spots from retrospective studies could yield comparable DNA methylome markers to those from prospectively collected neonatal blood samples. These results offer a means by which sample sizes of prediagnostic blood obtained from CC cases can be increased, especially that the rarity of such samples

in prospective settings represents a major roadblock. Therefore, follow-up studies seeking archived neonatal blood samples represent a promising way forward that could help the identification of epigenetic markers of CC risk using a hypothesis-free investigation.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-09671-3 OPEN

Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight

Leanne K. Küpers et al.#

Birthweight is associated with health outcomes across the life course, DNA methylation may be an underlying mechanism. In this meta-analysis of epigenome-wide association studies of 8,825 neonates from 24 birth cohorts in the Pregnancy And Childhood Epigenetics Consortium, we find that DNA methylation in neonatal blood is associated with birthweight at 914 sites, with a difference in birthweight ranging from −183 to 178 grams per 10% increase in methylation ($P_{Bonferroni} < 1.06 \times 10^{-7}$). In additional analyses in 7,278 participants, <1.3% of birthweight-associated differential methylation is also observed in childhood and adolescence, but not adulthood. Birthweight-related CpGs overlap with some Bonferroni-significant CpGs that were previously reported to be related to maternal smoking (55/914, $p = 6.12 \times 10^{-74}$) and BMI in pregnancy (3/914, $p = 1.13 \times 10^{-3}$), but not with those related to folate levels in pregnancy. Whether the associations that we observe are causal or explained by confounding or fetal growth influencing DNA methylation (i.e. reverse causality) requires further research.

Correspondence and requests for materials should be addressed to D.A.L. (email: d.a.Lawlor@bristol.ac.uk) or to C.L.R. (email: caroline.relton@bristol.ac.uk) or to H.S. (email: h.snieder@umcg.nl) or to J.F.F. (email: j.felix@erasmusmc.nl). #A full list of authors and their affiliations appears at the end of the paper.

Intrauterine exposures, such as maternal smoking, pre-pregnancy body mass index (BMI), hyperglycaemia, hypertension, folate and famine are associated with fetal growth and hence birthweight[1–6]. Observational studies show that birthweight is also associated with later-life health outcomes, including cardio-metabolic and mental health, some cancers and mortality[7–11]. In these long-term associations, birthweight may act as a proxy for potential effects of intrauterine exposures[12,13]. Several mechanisms may explain the associations of intrauterine exposures with birthweight and later-life health as we illustrate in Fig. 1. Our overall conceptual framework in this study was that the intrauterine environment induces epigenetic alterations, which influence fetal growth and hence correlate with birthweight. This is partly supported by previous large-scale epigenome-wide association studies (EWAS) that have reported associations of relevant maternal pregnancy exposures, including smoking, air pollution and BMI, with DNA methylation in offspring neonatal blood[14–16]. However, whilst four previous EWAS have observed associations of DNA methylation with birthweight[17–20], the evidence to date has been limited in scale and power with sample sizes ranging from approximately 200 to 1000.

In this study, we hypothesised that there are associations between DNA methylation and birthweight. We further aimed to explore if these epigenetic alterations are associated with later disease outcomes (Fig. 1). If birthweight is a proxy for a range of adverse prenatal exposures, we might expect neonatal blood DNA methylation to be associated with birthweight. However, we acknowledge that any associations of DNA methylation with birthweight may be explained by confounding[21] or reflect fetal growth influencing DNA methylation.

Here we present a large meta-analysis of multiple EWAS to explore associations between neonatal blood DNA methylation and birthweight. In further analyses, we explore whether any birthweight-associated differential methylation persists at older ages. To aid functional interpretation, we (i) explore the overlap of identified cytosine-phosphate-guanine sites (CpGs) that are differentially methylated in relation to birthweight with those known to be associated with intrauterine exposure to smoking, famine and different levels of BMI and folate; (ii) associate DNA methylation at identified CpGs with gene expression and (iii) explore potential causal links with birthweight and later-life health using Mendelian randomization (MR)[22]. We show that DNA methylation in neonatal blood is associated with birthweight and some of the differential methylation is also observed in childhood and adolescence, but not in adulthood. Also, we show overlap between birthweight-related CpGs and CpGs related to intrauterine exposures. Potential causality of the associations needs to be studied further.

Results

Participants. We used data from 8825 neonates from 24 studies in the Pregnancy And Childhood Epigenetics (PACE) Consortium, representing mainly European, but also African and Hispanic ethnicities with similar proportions of males and females. Details of participants used in all analyses are presented in Table 1, Supplementary Data 1 and study-specific Supplementary Methods.

Meta-analysis. Primary, secondary and follow-up analyses are outlined in the study design in Fig. 2. Methylation at 8170 CpGs, measured in neonatal blood using the Illumina Infinium® HumanMethylation450 BeadChip assay and adjusted for cell-type heterogeneity[23–25], was associated with birthweight (false discovery rate (FDR) <0.05), of which 1029 located in or near 807 genes survived the more stringent Bonferroni correction ($p < 1.06 \times 10^{-7}$, Supplementary Data 2). We observed both positive (45%) and negative (55%) directions of associations between methylation levels of these 1029 CpGs and birthweight (Fig. 3) and these CpGs were spread throughout the genome (orange track (1) in Fig. 4 and Supplementary Fig. 1). We found evidence of between-study heterogeneity ($I^2 >$ 50%) for 115 of the 1029 sites (Supplementary Data 2), thus we prioritised 914 CpGs, located in or near 729 genes, based on $p < 1.06 \times 10^{-7}$ and $I^2 \leq 50\%$ for further analyses (Fig. 3 and orange track (1) in Fig. 4). The CpG with the largest positive association was cg06378491 (in the gene body of *MAP4K2*). For each 10% increase in methylation at this site, birthweight was 178 g higher (95% confidence interval (CI): 138, 218 g). The CpG with the largest negative association was cg10073091 (in the gene body of *DHCR24*), which showed a 183 g decrease in birthweight per 10% increase in methylation (95% CI: −225, −142 g). The CpG with the smallest *P*-value and $I^2 \leq 50\%$ was cg17714703 (in the gene body of *UHRF1*), which showed a 130 g increase in birthweight for 10% increase in methylation (95% CI: 109, 151 g).

Findings were consistent with results from our main analyses when restricted to participants of European ethnicity, with a Pearson correlation coefficient for effect estimates of 0.99 for the 914 birthweight-associated CpGs (Supplementary Fig. 2, blue track (2) in Fig. 4 and Supplementary Data 3) and 0.90 for all 450k CpGs. Comparing the main meta-analysis to the four Hispanic cohorts and

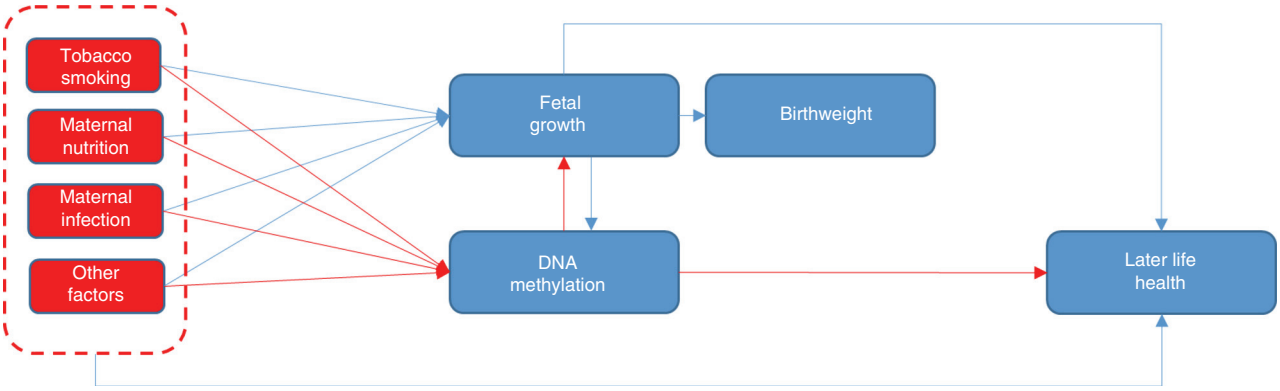


Fig. 1 Hypothetical paths that might link intrauterine exposures to DNA methylation, birthweight and later-life health outcomes. Red arrows summarise the paths that have motivated the analyses undertaken in this study (i.e. that maternal environmental exposures influence DNA methylation that in turn influences fetal growth and hence birthweight). The EWAS meta-analysis undertaken sought to identify methylation associated with birthweight. Blue arrows summarise other plausible paths, including that maternal exposures influence fetal growth first and it then influences DNA methylation or that maternal exposures may influence fetal growth/birthweight and later-life health outcomes through other pathways than DNA methylation

Table 1 Characteristics for the participating studies in the main meta-analysis for the association between neonatal blood DNA methylation and birthweight

Study	Total *N*	Normal birthweight, *N* (%)	High birthweight, *N* (%)	Birthweight (g)	Gestational age (wk)	Ethnicity	Boys, *N* (%)
ALSPAC	633	547 (86.4)	79 (12.5)	3512 ± 443	39.7 ± 1.3	European	301 (47.6)
CBC[a]: Hispanic	127	106 (83.5)	19 (15.0)	3445 ± 484	39.8 ± 1.3	Hispanic	74 (58.3)
CBC[a]: Caucasian	136	108 (79.4)	26 (19.1)	3625 ± 472	39.7 ± 1.5	European	79 (58.1)
CHAMACOS	283	236 (83.4)	44 (15.5)	3520 ± 446	39.3 ± 1.2	Hispanic	142 (50.1)
CHS[a]	199	168 (84.4)	28 (14.1)	3486 ± 476	40.2 ± 1.2	Mixed	79 (39.7)
EARLI	131	113 (86.3)	16 (12.2)	3507 ± 480	39.3 ± 1.0	Mixed	70 (53.4)
EXPOsOMICS: Rhea, Environage and Piccolipiu	324	297 (91.7)	22 (6.8)	3368 ± 437	39.4 ± 1.2	European	169 (52.1)
GECKO	255	206 (80.8)	46 (18.0)	3543 ± 533	39.7 ± 1.3	European	136 (53.3)
Gen3G	162	145 (89.5)	15 (9.3)	3408 ± 431	39.5 ± 1.1	European	74 (45.7)
Generation R	717	589 (82.1)	122 (17.0)	3572 ± 465	40.2 ± 1.1	European	372 (51.9)
GOYA[b]	947	649 (68.5)	294 (31.0)	3750 ± 501	40.4 ± 1.3	European	483 (51.0)
Healthy Start: African American	77	-	-	3059 ± 358	38.9 ± 1.3	African American	42 (54.5)
Healthy Start: Hispanic	115	-	-	3322 ± 395	39.1 ± 1.1	Hispanic	55 (47.8)
Healthy Start: Caucasian	240	220 (91.7)	14 (5.8)	3325 ± 425	39.3 ± 1.1	European	125 (52.1)
INMA	166	-	-	3297 ± 400	39.9 ± 1.2	European	82 (49.4)
IOW F2	118	97 (82.2)	17 (14.4)	3432 ± 525	39.7 ± 1.6	European	59 (50.0)
MoBa1	1066	795 (74.6)	251 (23.5)	3644 ± 544	39.5 ± 1.6	European	568 (53.3)
MoBa2	587	435 (74.1)	146 (24.9)	3701 ± 487	40.1 ± 1.2	European	329 (56.0)
MoBa3	205	153 (74.6)	51 (24.9)	3706 ± 491	39.8 ± 1.2	European	106 (51.7)
NCL[a]	792	592 (74.7)	192 (24.2)	3671 ± 506	40.0 ± 1.3	European	453 (57.2)
NEST: African American	99	-	-	3197 ± 534	39.3 ± 1.2	African American	47 (47.5)
NEST: Caucasian	111	94 (84.7)	13 (11.7)	3446 ± 471	39.5 ± 1.2	European	50 (45.0)
NHBCS	96	84 (87.5)	12 (12.5)	3509 ± 453	39.6 ± 1.1	European	53 (55.2)
PREDO	540	428 (79.3)	99 (18.3)	3572 ± 478	40.1 ± 1.2	European	264 (48.8)
PRISM	138	-	-	3385 ± 441	39.5 ± 1.1	Mixed	76 (55.1)
PROGRESS	143	-	-	3124 ± 387	38.6 ± 1.1	Hispanic	77 (53.8)
RICHS	89	52 (58.4)	23 (25.8)	3335 ± 734	38.9 ± 1.2	European	35 (39.3)
Project Viva	329	263 (79.9)	64 (19.5)	3623 ± 473	40.0 ± 1.2	European	168 (51.1)
Total *N*	8825	6377	1593				

Results are presented as mean ± SD or *N* (%). Normal birthweight: 2500−4000 g, high birthweight: >4000 g, low birthweight: <2500 g. Studies with mixed ethnicities analysed all participants together with adjustment for ethnicities. g: grams, wk: weeks, y: years. Full study names can be found in study-specific Supplementary Methods. For some studies the sample size for defining normal/high BW was too small

[a]CBC, CHS and NCL used heel prick blood spot samples instead of cord blood

[b]GOYA is a case-cohort study (cases are mothers with BMI>32 and controls are mothers randomly sampled from the underlying study population in which the cases were identified), in analyses where we included a random sample with a normal BMI distribution results were essentially the same as in the main analyses

the two African cohorts revealed that 94.9% and 74.0% of the 914 CpGs showed consistent direction of association, with Pearson correlation coefficients for point estimates of 0.82 and 0.48, respectively (Supplementary Data 3). In leave-one-out analyses, in which we reran the main meta-analysis repeatedly with one of the 24 studies removed each time, there was no strong evidence that any one study influenced findings consistently across the 914 differentially methylated CpGs that passed Bonferroni correction and for which between-study heterogeneity had an $I^2 \leq 50\%$. For 139/914 CpGs (15.2%) the difference in mean birthweight for a 10% greater methylation at that site varied by ≥20% with removal of a study, but the study resulting in the change was different for different CpGs. Supplementary Fig. 3.1-3.20 show the results for a random 10 plots where removal of one study changed the result by 20% or more and a random 10 where this was not the case; full results are available on request from the authors. Findings were broadly consistent when birthweight was categorised to high (>4000 g, $n = 1593$) versus normal (2500–4000 g, $n = 6377$) (Supplementary Data 4, yellow track (5) in Fig. 4) and when we did not exclude neonates born preterm or to women with pre-eclampsia or diabetes (Supplementary Fig. 4 and Supplementary Data 5A and 5C, and red track (3) in Fig. 4). Without these exclusions, we were able to examine associations with low (<2500 g, $n = 178$) versus normal (2500–4000 g, $n = 4197$) birthweight, though statistical power was still limited. Four CpGs were associated with low versus normal birthweight (Bonferroni-corrected threshold), none of which overlapped with the 914 CpGs from the main analysis (Supplementary Data 5B, purple track (4) in Fig. 4). We identified that 161 of the 914 differentially methylated CpGs potentially contained a single-nucleotide polymorphism (SNP) at cytosine or guanine positions (i.e. polymorphic CpGs; Supplementary Data 6). Polymorphic CpGs may affect probe binding and hence measured DNA methylation levels[26,27]. We used one of the largest studies (ALSPAC; $n = 633$) to explore this. We found no indication of bimodal distributions for any of the 161 CpGs suggesting SNPs had not markedly affected methylation measurements at these sites (dip test *p*-values: 0.299–1.00)[28–30].

Analyses at older ages. We took the 914 neonatal blood CpGs that were associated with birthweight at Bonferroni-corrected statistical significance and with $I^2 \leq 50\%$ and examined their associations with birthweight when measured in blood taken in childhood (2–13 years; $n = 2756$ from 10 studies), adolescence (16–18 years; $n = 2906$ from six studies) and adulthood (30–45 years; $n = 1616$ from three studies). Only participants from ALSPAC, CHAMACOS and Generation R had also contributed to the main neonatal blood EWAS. In childhood, adolescence and adulthood, we observed 87, 49 and 42 of the 914 CpGs to be nominally associated with birthweight ($p < 0.05$).

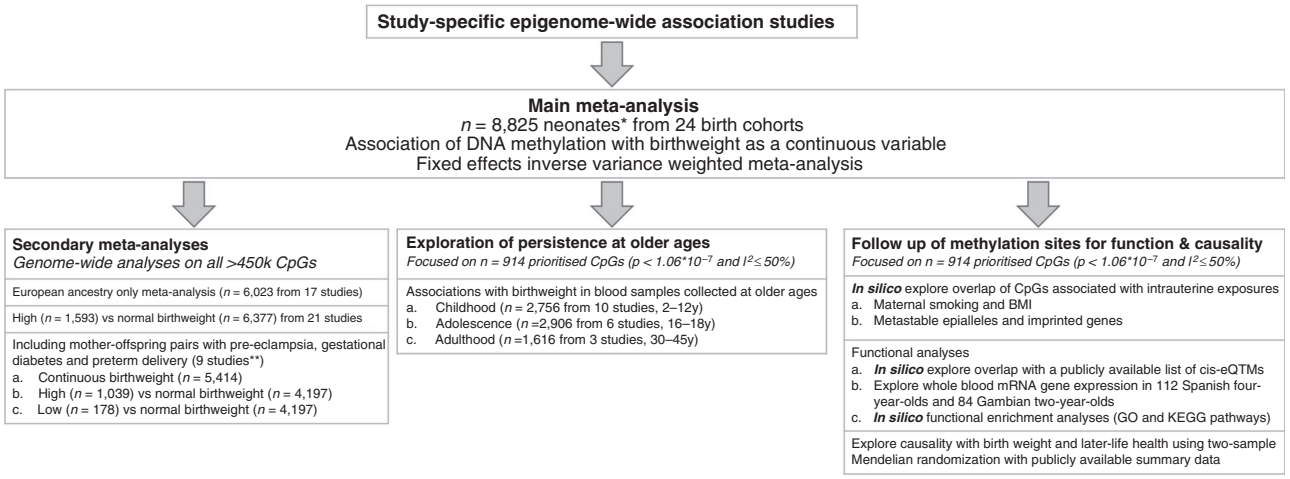


Fig. 2 Design of the study. Schematic representation of the main meta-analysis, secondary meta-analyses, follow-up analyses and exploration of persistence at older ages. *We removed multiple births from all analyses and excluded preterm births (<37 weeks) and offspring of mothers with pre-eclampsia or diabetes (three major pathological causes of differences). **For sufficient power in the low vs normal BW analyses, we only included nine studies with >10 low birthweight cases

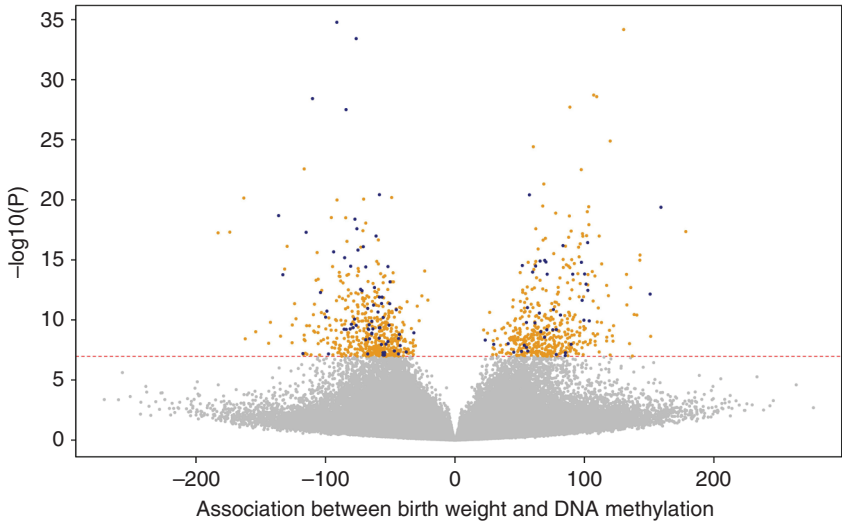


Fig. 3 Volcano plot showing the direction of associations of DNA methylation with birthweight in 8825 neonates from 24 studies. The *X*-axis represents the difference in birthweight in grams per 10% methylation difference, the *Y*-axis represents the $-\log_{10}(P)$. The red line shows the Bonferroni-corrected significance threshold for multiple testing ($p < 1.06 \times 10^{-7}$). Highlighted in orange are the 914 CpGs with $p < 1.06 \times 10^{-7}$ and $I^2 \leq 50\%$ and highlighted in blue are the 115 CpGs with $p < 1.06 \times 10^{-7}$ and $I^2 > 50\%$

All these CpGs showed consistent directions of association. Ten CpGs showed differential methylation across all four age periods. However, only a minority survived Bonferroni correction for 914 tests ($p < 5.5 \times 10^{-5}$): 12 (1.3%), 1 (0.1%) and 0 CpGs in childhood, adolescence and adulthood, respectively (Supplementary Data 7; the 12 CpGs that persisted in childhood are presented in the green track (6) in Fig. 4). Of the 914 CpGs, 50, 52 and 49% showed consistency in direction of association in childhood, adolescence and adulthood, but correlations of the associations of DNA methylation and birthweight between methylation measured in infancy and that measured in childhood, adolescence and adulthood were weak (Pearson correlation coefficients: 0.15, 0.06 and 0.02, respectively).

Intrauterine factors. We observed enrichment of previously published maternal smoking-related CpGs in the birthweight-associated CpGs[14] (55/914 (6.0%) $p_{\text{enrichment}} = 6.12 \times 10^{-74}$, of which cg00253658 and cg26681628 also showed persistent methylation differences in the look-up in childhood). We additionally found enrichment of maternal BMI-related CpGs in the list of birthweight-related CpGs[15] (3/914 (0.3%) $p_{\text{enrichment}} = 1.13 \times 10^{-3}$). All directions of association were consistent with the birthweight-lowering influence of maternal smoking or the positive association of maternal BMI with birthweight (Supplementary Data 8). We did not find evidence for overlap with plasma folate[31]. For famine, we were unable to explore overlap with DNA methylation at the Bonferroni-significant level as the previous EWAS of famine only reported results that reached a FDR level of statistical significance[32]. In additional analyses for overlap between all FDR hits from the birthweight EWAS with those FDR hits presented in the smoking, maternal BMI, folate and famine EWAS, we found an overlap of 430/8170 CpGs (5.3%, $p_{\text{enrichment}} = 7.38 \times 10^{-132}$) for smoking, 584/8170 CpGs (7.1%, $p_{\text{enrichment}} = 3.34 \times 10^{-62}$) for maternal BMI and 14/8170 (0.2%, $p_{\text{enrichment}} = 0.02$) for folate. For famine we did not observe overlap.

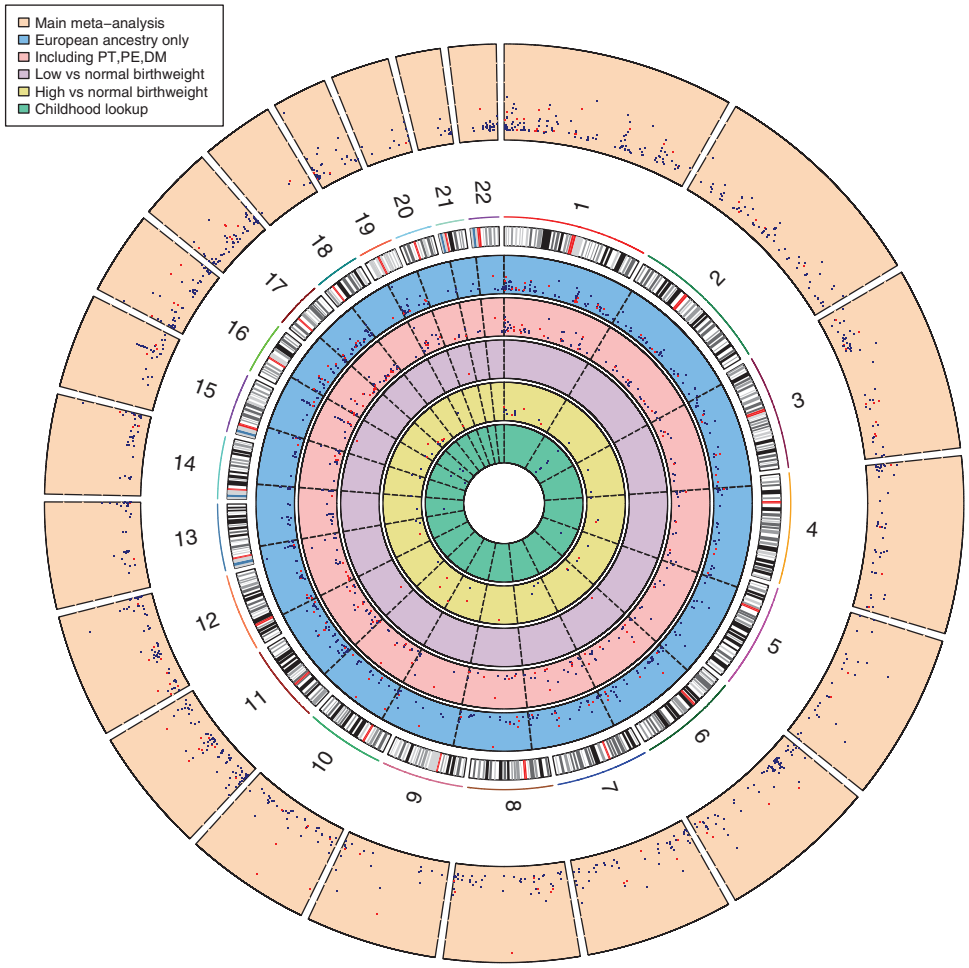


Fig. 4 Circos plot showing the (Bonferroni-corrected $p < 1.06 \times 10^{-7}$) results for associations of DNA methylation with birthweight. Results are presented as CpG-specific associations ($-\log_{10}(P)$, each dot represents a CpG) by genomic position, per chromosome. From outer to inner track: [1, orange] Main analysis results for associations between DNA methylation and birthweight as a continuous measure ($n = 8825$), [2, blue] Results from participants from European ethnicity only, DNA methylation and birthweight as a continuous measure ($n = 6023$), [3, red] Results from analysis without exclusion for preterm births, pre-eclampsia and maternal diabetes, DNA methylation and birthweight as a continuous measure $n = 5414$), [4, purple] Results from logistic regression analysis without exclusion for preterm births, pre-eclampsia and maternal diabetes, for low ($n = 178$) vs normal ($n = 4197$) birthweight, [5, yellow] Results from logistic regression analysis for associations between DNA methylation and high ($n = 1590$) vs normal ($n = 6114$) birthweight, [6, green] Results from look-up analysis in methylation samples taken during childhood and its association with birthweight as a continuous measure ($n = 2756$). Track 1: highlighted in red are 115 CpGs with $I^2 > 50\%$. Tracks 2–6: highlighted in red are CpGs that were not found in the 914 main meta-analysis hits (though note differences in sample size and hence statistical power for different analyses presented in the different tracks)

Metastable epialleles and imprinted genes. We tested the birthweight-associated CpGs for enrichment of metastable epialleles (loci for which the methylation state is established in the peri-conceptional period[33,34]). We additionally tested for enrichment of CpGs annotated to imprinted genes (loci that depend on the maintenance of parental-origin-specific methylation marks in the pre-implantation embryo, some of which are known to regulate fetal growth[35,36]). We did not find evidence of enrichment for metastable epialleles (3/1936 metastable epialleles overlap a birthweight-associated CpG), imprinting control regions (0/741) or imprinted gene transcription start sites (5/1728) (Supplementary Data 9).

Comparison with GWAS for birthweight. To compare these EWAS results to those from genetic studies, we used the 60 recently published fetal SNPs associated with birthweight in a GWAS meta-analysis of 153,781 newborns[37] and mapped the CpG sites identified in the EWAS to these SNPs to seek evidence of co-localisation of genetic and epigenetic variation (Supplementary Data 10). We repeated this for the 10 recently published maternal SNPs associated with birthweight in a GWAS meta-analysis of 86,577 women[38] (Supplementary Data 11). We observed that one or more of the 914 birthweight-associated CpGs were within +/−2Mb of 34/60 fetal and all 10 maternal birthweight-associated SNPs. Of the 34 fetal SNPs, three were located in the same gene as the CpG, as was one of the ten maternal SNPs. Ten fetal and four maternal SNPs were within 100 kb of identified CpGs. In a look-up of the fetal and maternal SNPs from GWAS of birthweight in an online cord blood methylation quantitative trait loci (mQTL) database (mqtldb.org[39]), 35 fetal and four maternal SNPs affected methylation at some CpG(s), but none at the 914 birthweight-associated CpGs specifically.

Functional analyses. We compared the 914 birthweight-related CpGs with a recently published list of 18,881 expression quantitative trait methylation sites (cis-eQTMs, +/−250 kb around

the transcription start site), CpG sites known to correlate with gene expression, from whole blood samples of 2101 Dutch adult individuals. We found that 82 of the 914 birthweight-associated CpGs were associated with gene expression of 98 probes (cis-eQTMs)[40] ($p_{enrichment} < 1.73 \times 10^{-11}$, Supplementary Data 12). Additionally, in 112 Spanish 4-year-olds[41], we observed that 19 CpGs were inversely associated with whole blood mRNA gene expression and four CpGs were positively associated with gene expression (FDR<0.05, Supplementary Data 13). Of these 23 CpGs, 13 were also found in the publicly available cis-eQTM list[40]. In 84 Gambian children (age 2 years)[42], we found two CpGs that were inversely associated with whole blood mRNA gene expression, but neither were found in the Spanish results or the publicly available cis-eQTM list. The 914 birthweight-associated CpGs showed no functional enrichment of Gene Ontology (GO) terms or Kyoto Encyclopedia of Genes and Genomes (KEGG) terms (FDR<0.05).

Mendelian randomization. We aimed to explore causality using MR analysis, in which genetic variants associated with methylation levels (methylation quantitative trait loci (mQTLs)) are used as instrumental variables to appraise causality. For 788 (86%) of the 914 birthweight-associated CpGs, no mQTLs were identified in a publicly available mQTL database[39]. For 108 (86%) of the remaining 126 CpGs, only one mQTL was identified and for the remainder none had more than four mQTLs (Supplementary Data 14 provides a complete list of all mQTLs identified for these 126 CpGs). Many of the currently available methods that can be used as sensitivity analyses to explore whether MR results are biased by horizontal pleiotropy (a single mQTL influencing multiple traits) require more than one genetic instrument (here mQTLs) and even with two or three this can be difficult to interpret[43]. Having determined that it was not possible to undertake MR analyses of 86% of the birthweight-related differentially methylated CpGs (because we did not identify any mQTLs), and for the majority of the remaining CpGs we would not have been reliably able to distinguish causality from horizontal pleiotropy (because only one mQTL could be identified), we decided not to pursue MR analyses further.

Discussion

This large-scale meta-analysis shows that birthweight is associated with widespread differences in DNA methylation. We observed some enrichment of birthweight-associated CpGs among sites that have previously been linked to smoking during pregnancy[14] and pre-pregnancy BMI[15], consistent with the hypothesis that epigenetic pathways may underlie the observational associations of those prenatal exposures with birthweight[21,44,45]. However, the actual overlap in this analysis was modest, likely explained by the adjustments for maternal smoking and BMI in the EWAS analyses. The overlap that we observed with pregnancy smoking-related CpGs may reflect the possibility that smoking-related CpGs capture smoking better than self-report[46,47], in line with expectations of pregnant women underreporting their smoking behaviour. Adjustment for maternal smoking and BMI may have masked a greater level of overlap between our results and EWAS of these two maternal exposures. The fact that we find an association of DNA methylation across the genome with birthweight provides some support for our conceptual framework shown in Fig. 1. However, we acknowledge that the associations that we have observed may also be explained by causal effects of maternal pregnancy exposures on both DNA methylation and fetal growth, as well as subtle inflammatory responses in cell-type proportions associated with maternal smoking that might not have been completely captured with the currently available cell type estimation methods.

The differential methylation associated with birthweight in neonates persisted only minimally across childhood and into adulthood. Larger (preferably longitudinal) studies are needed to explore persistent differential methylation in more detail and with better power at older ages. It is possible that inclusion of the Gambia study in the childhood EWAS (which was the only non-European study in these analyses and was not included in the main meta-analyses with neonatal blood) might have impacted these results, although this study made up just 7% of the total child follow-up sample. A rapid attenuation of differential methylation in relation to birthweight in the first years after birth has previously been reported[19], but our sample size for these analyses may have been too small to detect persistence. This rapid decrease, if real, may indicate a reduction in the dose of the child's exposure to maternal factors such as smoking once the offspring is delivered, with that reduction continuing as the child ages. Persistence of birthweight-related differential DNA methylation may not necessarily be a prerequisite for long-term effects, as transient differential methylation in early life may cause lasting functional alterations in organ structure and function that predispose to later adverse health effects.

Methylation is known to be associated with gene expression[48]. However, we found no consistent associations between birthweight-related methylation and gene expression in two childhood studies. This could be due to the relatively small sample sizes, differences in ethnicities, age, or platforms to measure gene expression. The use of blood, which is likely only a possible surrogate tissue for fetal growth phenotypes, for gene expression analysis might also explain the lack of findings. We did find multiple cis-eQTMs among the birthweight-related CpGs at which methylation was related to gene expression in blood when using a publicly available database from a larger adult sample[40], providing some evidence that birthweight-related differentially methylated CpGs may be associated with gene expression. These initial in silico association analyses need further exploration to establish any underlying causal mechanisms.

In observational studies, birthweight has repeatedly been associated with a range of later-life diseases. Change in DNA methylation has been hypothesized as a potential mechanism linking early exposures, birthweight and later health (Fig. 1). We originally aimed to explore this using MR analysis. For the vast majority of the birthweight-associated CpGs, no genetic instrumental variables were available. For the remaining 126 CpGs, only one mQTL was available, which would make it impossible to disentangle causality from horizontal pleiotropy. To ensure a strong basis for future MR analyses on this topic, there is a clear need for a more extensive mQTL resource.

Strengths of this study are its large sample size and the extensive analyses that we have undertaken. In a post hoc power calculation based on the sample size of 8825 with a weighted mean birthweight of 3560 g (weighted mean standard deviation (SD): 483 g) and with an alpha set at the Bonferroni-corrected level of $P < 1.06 \times 10^{-7}$ we had 80% power, with a two-sided test, to detect a minimum difference of 0.13 SD (63 g) in birthweight for each SD increase in methylation. The difference in methylation corresponding to a 1 SD increase differs per CpG, as it depends on the distribution of the methylation values. We acknowledge that smaller differences which might be clinically or biologically relevant may not have been identified in the current analysis. Nonetheless, to our knowledge this analysis has brought together all studies currently available with relevant data and is the largest published study of this association. DNA methylation patterns in neonatal blood, whilst easily accessible in large numbers, may not reflect the key tissue of importance in relation to birthweight. DNA methylation and gene expression in placental tissue may be important targets for future studies. DNA methylation varies between leucocyte subtypes[49] and we used an adult whole blood reference to correct for this in the main analyses[23,24], as the study-specific analyses were completed before the widespread availability of specific cord blood reference datasets[50,51]. However, we observed very similar findings in two studies

(Generation R and GECKO) when we compared the results with those using one of the currently available cord blood references[50]. Although we adjusted for potential major confounders that may affect both methylation and fetal growth, we acknowledge that the main results cannot ascertain causality. That is, whilst we have hypothesised that variation in fetal DNA methylation influences fetal growth and hence birthweight, and undertaken the analyses accordingly, we cannot exclude the possibility that differences in neonatal blood DNA methylation are caused by variation in fetal growth itself, or that the association is confounded by factors, including maternal smoking and BMI, that independently influence both fetal growth and DNA methylation (as suggested in Fig. 1). The 450k array that was used to measure genome-wide DNA methylation only covers 1.7% of the total number of CpGs present in the genome and specifically targets CpGs in promoter regions and gene bodies[52]. We removed the CpGs that were flagged as potentially cross-reactive, as the measured methylation levels may represent methylation at either of the potential loci. Also, although we did not find evidence for polymorphic effects for the 161 potentially polymorphic CpGs in ALSPAC, we cannot completely exclude these potential polymorphic effects in the meta-analysed results. The majority of participants were of European ethnicity and when analyses were restricted to those of European ethnicity the results were essentially identical to those with all studies included. Direct comparisons of the main analysis with analyses in those of Hispanic or of African ethnicity for the 914 hits suggested strong correlations with Hispanic but weaker with African ethnicity. However, these results need to be treated with caution. First, we had very few studies of Hispanic and African populations. Second, we only compared the initial hits from the main meta-analysis with all ethnicities included. A detailed exploration of ethnic differences would require similar large samples for each ethnic group and within ethnic EWAS, which is beyond the scope of the data currently available.

Neonatal blood DNA methylation at many sites across the genome is associated with birthweight. Further research is required to determine if these are causal and if so whether they mediate any long-term effect of intrauterine exposures on future health.

Methods

Participants. In the main EWAS meta-analysis we explored associations of neonatal blood DNA methylation with birthweight using data from 8825 neonates from 24 studies in the PACE Consortium[53] (Table 1). We removed multiple births from all analyses and excluded preterm births (<37 weeks) and offspring of mothers with pre-eclampsia or diabetes (three major pathological causes of differences in fetal growth). In follow-up analyses, we explored whether any sites found in the main analysis were discernible in relation to birthweight when examined in DNA from blood drawn during childhood (2–13 years; 2756 children from 10 studies), adolescence (16–18 years; 2906 adolescents from six studies) or adulthood (30–45 years; 1616 adults from three studies), see Supplementary Data 1B. Informed consent was obtained from all participants, and all studies received approval from local ethics committees. Study-specific methods and ethical approval statements are provided in Supplementary Methods.

Birthweight, DNA methylation and covariates. Our primary outcome was birthweight on a continuous scale (grams), adjusted for gestational age, and measured immediately after birth or retrospectively reported by mothers in questionnaires. In secondary analyses, we categorised and compared associations with high (>4000 g, $n = 1593$) versus normal (2500–4000g, $n = 6377$) birthweight. We also explored all associations with (continuous and categorical) birthweight in analyses that did not exclude women with pre-eclampsia, diabetes or preterm delivery, which also resulted in enough cases to explore low (<2500 g, $n = 178$) versus normal (2500–4000 g, $n = 4197$) birthweight (Supplementary Data 1C shows the characteristics of participants). Primary, secondary and follow-up analyses are outlined in the study design in Fig. 2. DNA methylation was measured in neonatal blood samples using the Illumina Infinium® HumanMethylation450 BeadChip assay. All participants had cord blood samples except for three studies with heel stick blood spots ($n = 1254$ [14.2%]). After study-specific laboratory analyses, quality control, normalisation, and removal of control probes ($n = 65$) and probes that mapped to the X ($n = 11,232$) and Y ($n = 370$) chromosomes, we included 473,864 CpGs. DNA methylation is expressed as the proportion of cells in which the DNA was methylated at a specific site and hence takes values from zero to one. We converted this to a percentage and present differences in mean birthweight per 10% higher DNA methylation level at each CpG. All analyses were adjusted for gestational age at delivery, child sex, maternal age at delivery, parity (0/≥1), smoking during pregnancy (no smoking/stopped in early pregnancy/smoking throughout pregnancy), pre-pregnancy BMI, socio-economic position, technical variation, and estimated white blood cell proportions (B-cells, CD8+ T-cells, CD4+ T-cells, granulocytes, NK-cells and monocytes)[23–25]. In studies with participants from multiple ethnic groups, each group was analysed separately and results were added to the meta-analyses as separate studies. Further details are provided in the study-specific Supplementary Methods.

Statistical methods. Robust linear (birthweight as a continuous outcome) or logit (binary birthweight outcomes) regression EWAS were undertaken within each study according to a pre-specified analysis plan. Quality control, normalisation and regression analyses were conducted independently by each study. After confirming comparability of study-specific summary statistics[54], we combined results using a fixed effects inverse variance weighted meta-analysis[55]. The meta-analysis was done independently by two study groups and the results were compared in order to minimise the likelihood of human error. We show (two-sided) results after correcting for multiple testing using both the FDR<0.05[56] and the Bonferroni correction ($p < 1.06 \times 10^{-7}$). We completed follow-up analyses for differentially methylated CpGs that reached the Bonferroni-adjusted threshold and did not show large between-study heterogeneity[57] ($I^2 \leq 50\%$). We annotated the nearest gene for each CpG using the UCSC Genome Browser build hg19[58,59]. We explored whether between-study heterogeneity might be explained by differences in ethnicity between studies, by repeating the meta-analysis including only participants of European ethnicity, which was by far the largest ethnic subgroup ($n = 6023$ from 17 studies) (Fig. 2). Ethnicity was defined using maternal or self-report, unless specified otherwise in study-specific Supplementary Methods. We also did meta-analyses only including the Hispanic studies and only including the African American studies and present those results for illustrative purposes only, given the much smaller sample size. All analyses were performed using R[60], except for the meta-analysis which was performed using METAL[55]. We removed CpGs that co-hybridised to alternate sequences (i.e. cross-reactive sites), because we cannot distinguish whether the differential methylation is at the locus that we have reported or at the one that the probe cross-reacts with. We compared the birthweight-related CpGs to lists of CpGs that are potentially influenced by a SNP (polymorphic sites)[26,27]. For these CpGs, we determined if DNA methylation levels were influenced by nearby SNPs, by assessing whether their distributions deviated from unimodality using Hartigans' dip test[28,29] and visual inspection of density plots in $n = 742$ cord blood samples in the ALSPAC study.

Analyses at older ages. Analyses of the associations with DNA methylation in blood collected in childhood, adolescence and adulthood followed the same covariable adjustment and methods as for the main analyses ($p < 5.5 \times 10^{-5}$ for 914 tests). All participants and studies in these analyses at older ages had not been included in the main meta-analysis in neonatal blood, except for ALSPAC ($n = 633$ in neonatal analyses, $n = 605$ in childhood and $n = 526$ in adolescence), CHAMACOS ($n = 283$ in neonatal analyses and $n = 191$ in childhood) and Generation R ($n = 717$ in neonatal analyses and $n = 372$ in childhood). Characteristics are shown in study-specific Supplementary Methods and Supplementary Data 1B.

Intrauterine factors. We used a hypergeometric test to explore the extent to which any of the birthweight-related CpGs overlapped with those previously associated with intrauterine exposure to smoking[14] ($n = 568$ CpGs), BMI[15] ($n = 104$ CpGs) and plasma folate[31] ($n = 48$ CpGs), using the same (Bonferroni-corrected) cut-off for statistical significance. No CpGs reached the Bonferroni-corrected cut-off for famine[32]. We additionally appraised this overlap using the FDR<0.05 cut-off for all traits ($n = 8170$ birthweight-related CpGs, $n = 6703$ smoking-related CpGs, $n = 16,067$ BMI-related CpGs, $n = 443$ folate-related CpGs, $n = 7$ famine-related CpGs). These FDR results were available from the publications for smoking, folate and famine, and we obtained them from the corresponding author for BMI.

Metastable epialleles and imprinted genes. We tested the birthweight-associated CpGs for enrichment of metastable epialleles and CpGs associated with imprinted genes. The metastable epialleles were derived from a recently published study that identified 1936 putative metastable epialleles[34]. For imprinted genes, we first identified a set of CpGs falling within a curated set of imprinting control regions; differentially methylated regions controlling the parental-specific expression of one or more imprinted genes[36]. Second, we extracted the set of imprinting control region controlled genes from the above source and identified all 450k CpGs within +/−10kbp of the gene transcription start site, including all known alternative TSS identified in grch37.ensembl.org using biomaRt[61,62].

Comparison with GWAS for birthweight. We compared the birthweight-associated CpGs with the 60 SNPs from the most recent GWAS meta-analyses of fetal genotype associations with birthweight in >150,000 newborns[37] and with the 10 SNPs from the most recent GWAS meta-analysis of maternal genotype associations with birthweight in >86,000 women[38]. With this comparison we checked if the EWAS top hits were located within a 4 Mb window (+/− 2 Mb)

surrounding these SNPs. We additionally checked whether SNPs and CpGs were located in the same gene.

Functional analyses. To explore the association of methylation with gene expression, we compared birthweight-related CpGs with a recently published list of 18,881 cis-eQTMs from whole blood samples of 2101 Dutch adult individuals[40]. With a hypergeometric test, we calculated enrichment of cis-eQTMs in the list of birthweight-associated CpGs. We further explored methylation of birthweight-associated CpGs in relation to whole blood mRNA gene expression (transcript levels) within a 500 kb region of the CpGs (+/−250 kb, FDR<0.05) in 112 Spanish 4-year-olds[41] and 84 Gambian 2-year-olds[42] (Supplementary Methods). To better understand the potential mechanisms linking DNA methylation and birthweight, we explored the potential functions of the birthweight-associated CpGs using GO and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analyses. We used the missMethyl R package[63], which enabled us to correct for the number of probes per gene on the 450k array, based on the November 2018 version of the GO and KEGG source databases. To filter out the large, general pathways we set the number of genes for each gene set between 15 and 1000, respectively. We calculated FDR at 5% corrected *P*-values for enrichment.

Mendelian randomization. MR uses genetic variants as instrumental variables to study the causal effect of exposures on outcomes[64,65]. We aimed to use two-sample MR[22,66] to explore (a) evidence of a causal association of methylation levels at the identified CpGs with birthweight and (b) evidence of a causal association of these CpGs with later-life health outcomes (i.e. to explore our hypothesised causal mechanisms shown in Fig. 1). We did this by first searching a publicly available mQTL database[39] to identify cis-mQTLs within 1 Mb of each of the Bonferroni-corrected, with $I^2 \leq 50\%$, birthweight-related differentially methylated CpGs. These mQTLs could then be used as genetic instrumental variables for methylation levels of the birthweight-related CpGs. We then aimed to determine the association of these mQTLs with birthweight and later-life health outcomes from publicly available summary GWAS results[66].

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files or from the corresponding authors upon reasonable request. All summary statistics from this EWAS meta-analysis are available via doi: 10.5281/zenodo.2222287. A reporting summary for this Article is available as a Supplementary Information file.

Code availability

The code used for this EWAS meta-analysis is available from the authors upon request.

Received: 30 April 2018 Accepted: 18 February 2019
Published online: 23 April 2019

References

1. Tyrrell, J. et al. Genetic evidence for causal relationships between maternal obesity-related traits and birth weight. *JAMA* **315**, 1129 (2016).
2. Tyrrell, J. et al. Genetic variation in the 15q25 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) interacts with maternal self-reported smoking status during pregnancy to influence birth weight. *Hum. Mol. Genet.* **21**, 5344–5358 (2012).
3. Bakker, R., Steegers, E. A. P., Hofman, A. & Jaddoe, V. W. V. Blood pressure in different gestational trimesters, fetal growth, and the risk of adverse birth outcomes: the generation R study. *Am. J. Epidemiol.* **174**, 797–806 (2011).
4. Lawlor, D. A. et al. Association of existing diabetes, gestational diabetes and glycosuria in pregnancy with macrosomia and offspring body mass index, waist and fat mass in later childhood: Findings from a prospective pregnancy cohort. *Diabetologia* **53**, 89–97 (2010).
5. van Uitert, E. M. & Steegers-Theunissen, R. P. M. Influence of maternal folate status on human fetal growth parameters. *Mol. Nutr. Food. Res.* **57**, 582–595 (2013).
6. Painter, R. C., Roseboom, T. J. & Bleker, O. P. Prenatal exposure to the Dutch famine and disease in later life: an overview. *Reprod. Toxicol.* **20**, 345–352 (2005).
7. Whincup, P. H. et al. Birth weight and risk of type 2 diabetes a systematic review. *JAMA* **300**, 2886–2897 (2008).
8. Lawlor, D. A., Ronalds, G., Clark, H., Davey Smith, G. & Leon, D. A. Birth weight is inversely associated with incident coronary heart disease and stroke among individuals born in the 1950s: findings from the Aberdeen children of the 1950s prospective cohort study. *Circulation* **112**, 1414–1418 (2005).
9. O'Donnell, K. J. & Meaney, M. J. Fetal origins of mental health: the developmental origins of health and disease hypothesis. *Am. J. Psychiatry* **174**, 319–328 (2017).
10. McCormack, V. A., Silva, I. D. S., Koupil, I., Leon, D. A. & Lithell, H. O. Birth characteristics and adult cancer incidence: Swedish cohort of over 11,000 men and women. *Int. J. Cancer* **115**, 611–617 (2005).
11. Risnes, K. R. et al. Birthweight and mortality in adulthood: a systematic review and meta-analysis. *Int. J. Epidemiol.* **40**, 647–661 (2011).
12. Freathy, R. M. Can genetic evidence help us to understand the fetal origins of type 2 diabetes? *Diabetologia* **59**, 1850–1854 (2016).
13. Hanson, M. Birth weight and the fetal origins of adult disease. *Pediatr. Res.* **52**, 473–474 (2002).
14. Joubert, B. R. et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. *Am. J. Hum. Genet.* **98**, 680–696 (2016).
15. Sharp, G. C. et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. *Hum. Mol. Genet.* **26**, 4067–4085 (2017).
16. Gruzieva, O. et al. Epigenome-wide meta-analysis of methylation in children related to prenatal NO_2 air pollution exposure. *Environ. Health Perspect.* **125**, 104–110 (2017).
17. Adkins, R. M., Tylavsky, F. A. & Krushkal, J. Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight. *Chem. Biodivers.* **9**, 888–899 (2012).
18. Engel, S. M. et al. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. *Am. J. Epidemiol.* **179**, 834–842 (2014).
19. Simpkin, A. J. et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Hum. Mol. Genet.* **24**, 3752–3763 (2015).
20. Agha, G. et al. Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. *Clin. Epigenetics* **8**, 1–12 (2016).
21. Lawlor, D. A., Relton, C., Sattar, N. & Nelson, S. M. Maternal adiposity—a determinant of perinatal and offspring outcomes? *Nat. Rev. Endocrinol.* **8**, 679–688 (2012).
22. Lawlor, D. A. Commentary: Two-sample Mendelian randomization: opportunities and challenges. *Int. J. Epidemiol.* **45**, 908–915 (2016).
23. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 1–16 (2012).
24. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS. ONE.* **7**, e41361 (2012).
25. Aryee, M. J. et al. Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
26. Chen, Y. A. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics.* **8**, 203–209 (2013).
27. Naeem, H. et al. Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array. *BMC Genomics* **15**, 51 (2014).
28. Hartigan, J. & Hartigan, P. The dip test of unimodality. *Ann. Stat.* **13**, 70–84 (1985).
29. Maechler, M. Hartigan's dip test statistic for unimodality - corrected. R package. (2015).
30. Relton, C. L. et al. Data resource profile: accessible resource for integrated epigenomic studies (ARIES). *Int. J. Epidemiol.* **44**, 1181–1190 (2015).
31. Joubert, B. R. et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. *Nat. Commun.* **7**, 10577 (2016).
32. Tobi, E. W. et al. Early gestation as the critical time-window for changes in the prenatal environment to affect the adult human blood methylome. *Int. J. Epidemiol.* **44**, 1211–1223 (2015).
33. Rakyan, V. K., Blewitt, M. E., Druker, R., Preis, J. I. & Whitelaw, E. Metastable epialleles in mammals. *Trends Genet.* **18**, 348–351 (2002).
34. Van Baak, T. E. et al. Epigenetic supersimilarity of monozygotic twin pairs. *Genome. Biol.* **19**, 2 (2018).
35. Moore, G. E. et al. The role and interaction of imprinted genes in human fetal growth. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **370**, 20140074 (2015).
36. Monk, D. et al. Recommendations for a nomenclature system for reporting methylation aberrations in imprinted domains. *Epigenetics* 1–5 https://doi.org/10.1080/15592294.2016.1264561 (2016).
37. Horikoshi, M. et al. Genome-wide associations for birth weight and correlations with adult disease. *Nature* **4**, 1–20 (2016).
38. Beaumont, R. N. et al. Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. *Hum. Mol. Genet.* **27**, 742–756 (2018).

39. Gaunt, T. R. et al. Systematic identification of genetic influences on methylation across the human life course. *Genome. Biol.* **17**, 1–14 (2016).
40. Bonder, M. J. et al. Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2016).
41. Guxens, M. et al. Cohort profile: the INMA—INfancia y Medio Ambiente—(environment and childhood) project. *Int. J. Epidemiol.* **41**, 930–940 (2012).
42. Moore, S. E. et al. A randomized trial to investigate the effects of pre-natal and infant nutritional supplementation on infant immune development in rural Gambia: the ENID trial: early nutrition and immune development. *BMC Pregnancy Childbirth* **12**, 1–8 (2012).
43. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
44. Richmond, R. C., Timpson, N. J. & Sørensen, T. I. A. Exploring possible epigenetic mediation of early-life environmental exposures on adiposity and obesity development. *Int. J. Epidemiol.* **44**, 1191–1198 (2015).
45. Küpers, L. K. et al. DNA methylation mediates the effect of maternal smoking during pregnancy on birthweight of the offspring. *Int. J. Epidemiol.* **44**, 1224–1237 (2015).
46. Valeri, L. et al. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? *Epigenomics* **9**, 253–265 (2017).
47. Reese, S. E. et al. DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. *Env. Heal. Perspect.* https://doi.org/10.1289/EHP333 (2016).
48. Jaenisch, R. & Bird, A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat. Genet.* **33**(Suppl), 245–254 (2003).
49. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome. Biol.* **15**, R31 (2014).
50. Bakulski, K. M. et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. *Epigenetics.* **11**, 354–362 (2016).
51. Gervin, K. et al. Cell type specific DNA methylation in cord blood: a 450K-reference data set and cell count-based validation of estimated cell type composition. *Epigenetics.* **11**, 690–698 (2016).
52. Dedeurwaerder, S. et al. Evaluation of the infinium methylation 450K technology. *Epigenomics* https://doi.org/10.2217/epi.11.105 (2011).
53. Felix, J. F. et al. Cohort profile: pregnancy and childhood epigenetics (PACE) consortium. *Int. J. Epidemiol.* **16**, 10–14 (2017).
54. van der Most, P. J., Küpers, L. K., Snieder, H. & Nolte, I. QCEWAS: automated quality control of results of epigenome-wide association studies. *Bioinformatics* **33**, 1243–1245 (2017).
55. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
56. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat. Soc B* **57**, 289–300 (1995).
57. Higgins, J. P. T. & Thompson, S. G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).
58. Triche, T. FDb.InfiniumMethylation.hg19: annotation package for Illumina Infinium DNA methylation probes. R package version 2.2.0 (2014).
59. Carlson, M. & Maintainer, B. TxDb.Hsapiens.UCSC.hg19.knownGene: annotation package for TxDb object(s). R package version 3.2.2 (2015).
60. R Core Team. R: A language and environment for statistical computing. (Austria, 2013).
61. Durinck, S. et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
62. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
63. Phipson, B., Maksimovic, J. & Oshlack, A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics* **32**, 286–288 (2016).
64. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
65. Lawlor, D. A., Harbord, R. M., Sterne, J. A. C., Timpson, N. & Davey Smith, G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat. Med.* **27**, 1133–1163 (2008).
66. Hemani, G. et al. The MR-Base platform supports systematic causal inference across the human phenome. *eLife* **7**, e34408 (2018).

Acknowledgements

For all studies, acknowledgements can be found in Supplementary Information: Supplementary Acknowledgements. For all studies, funding statements can be found in Supplementary Information: Supplementary Funding.

Author contributions

L.K.K., D.A.L., C.L.R., H.S. and J.F.F. conceived and designed the study. Study-specific analyses were completed by G.C.S. (ALSPAC and GOYA), S.K.M. (BAMSE), R.R. (CBC), P.Y. (CHAMACOS), C.V.B. (CHS), K.M.B. (EARLI), A.G. and A.N. (EXPOsOMICS, The Gambia and MoBa3), S.A.S.L. (FLEHS1), L.K.K. (GECKO), C.A. (Gen3G), C.M. (Generation R), J.L. (Glaku), A.P.S. (Healthy Start), L.A.S. (INMA), F.I.R. (IOW F1), J.W.H. (IOW F2), D.A.V.D. P. (Lifelines), C.M.P. (MoBa1), S.E.R. (MoBa2), A.J.W. (NCL), D.D.J. (NEST), M.W. (NFBC66 and NFBC86), T.M.E. (NHBCS and RICHS), J.V.D. (NTR), C.J.X. (PIAMA), D.C. (PREDO), A.C.J. (PRISM), A.C.J. (PROGRESS), S.L. (Project Viva), R.C.H. (Raine), V.U. (STOPPA). L.K.K. and C.M. meta-analysed the results. L.K.K., C.M., G.C.S., P.Y., L.A.S., A.G., A.N. and M.J.S. performed follow-up analyses. L.K.K., D.A.L., C.L.R., H.S. and J.F.F. interpreted the results. L.K.K., with input from D.A.L., C.L.R., H.S. and J.F.F., wrote the first draft of the manuscript. All authors (L.K.K., C.M., G.C.S., P.Y., L.A.S., A.G., C.M.P., S.E.R., A.J.W., D.C., A.P.S., A.N., S.L., R.R., C.H., C.V.B., C.A., A.C.J., K.M.B., J.W.H., T.M.E., C.-J.X., R.-C.H., D.Avd.P., M.W., S.K.M., V.U., F.I.R., J.L., J.vD., S.A.S.L., T.G.R., M.C.M., E.A.N., Z.X., L.D., S.Z., W.Z., M.P., D.L.D., O.S., J.H.H., D.D.J., L.G., M.B., P.P., R.O.W., I.H-P., H.Z., M.R.K., U.G., C.J.M., L.J.B., J.M.V., M.-R.J., A.B., A.K.Ö., S.E., P.M.V., S.E.M., G.W., A.R.L.S., S.E.H., T.I.A.S., J.A.T., K.R., I.V.Y., K.K., T.S.N., M.J.S., Y.Y.G., L.R., M.K., A.A.L., B.E., K.H., H.M., R.L.M., T.D., M.V., L.B., A.A.B., L.A.C., W.K., D.A., M.dV., S.S., J.K., R.K., S.H.A., E.H., M.N.R., D.I.B., A.P.F., C.J.N., E.G., M.M., M.D.F., E.M., A.M.P., E.K., C.A., E.O., D.D., H.M.B., P.E.M., R.J.W., G.H.K., L.T., M.-F.H., J.S., M.C.M.-K., S.K.M., E.C., J.W., N.H., Z.H., E.B.B., G.D.S., V.W.V.J., R.T.L., W.N., S.J.L., D.A.L., C.L.R., H.S., J.F.F.) read and critically revised subsequent drafts.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-019-09671-3.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Journal peer review information: *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Leanne K. Küpers[1,2,3,4], Claire Monnereau[5,6,7], Gemma C. Sharp[1,8], Paul Yousefi[1,2,9], Lucas A. Salas[10,11], Akram Ghantous[12], Christian M. Page[13,14], Sarah E. Reese[15], Allen J. Wilcox[15], Darina Czamara[16], Anne P. Starling[17], Alexei Novoloaca[12], Samantha Lent[18], Ritu Roy[19,20], Cathrine Hoyo[21,22], Carrie V. Breton[23], Catherine Allard[24], Allan C. Just[25], Kelly M. Bakulski[26], John W. Holloway[27,28], Todd M. Everson[29], Cheng-Jian Xu[30,31], Rae-Chi Huang[32], Diana A. van der Plaat[33], Matthias Wielscher[34], Simon Kebede Merid[35], Vilhelmina Ullemar[36], Faisal I. Rezwan[28], Jari Lahti[37,38], Jenny van Dongen[39], Sabine A.S. Langie[40,41,42], Tom G. Richardson[1,2], Maria C. Magnus[1,2,13], Ellen A. Nohr[43], Zongli Xu[44], Liesbeth Duijts[4,44,45], Shanshan Zhao[46], Weiming Zhang[47], Michelle Plusquin[48,49], Dawn L. DeMeo[50], Olivia Solomon[8], Joosje H. Heimovaara[3], Dereje D. Jima[22,51], Lu Gao[23], Mariona Bustamante[11,52,53,54], Patrice Perron[24,55], Robert O. Wright[25], Irva Hertz-Picciotto[56], Hongmei Zhang[57], Margaret R. Karagas[10,58], Ulrike Gehring[59], Carmen J. Marsit[29], Lawrence J. Beilin[60], Judith M. Vonk[33], Marjo-Riitta Jarvelin[34,61,62,63], Anna Bergström[35,64], Anne K. Örtqvist[36], Susan Ewart[65], Pia M. Villa[66], Sophie E. Moore[67,68], Gonneke Willemsen[39], Arnout R.L. Standaert[40], Siri E. Håberg[13], Thorkild I.A. Sørensen[1,69,70], Jack A. Taylor[15], Katri Räikkönen[38], Ivana V. Yang[71], Katerina Kechris[45], Tim S. Nawrot[48,72], Matt J. Silver[67], Yun Yun Gong[73], Lorenzo Richiardi[74,75], Manolis Kogevinas[11,53,54,76], Augusto A. Litonjua[50], Brenda Eskenazi[9,77], Karen Huen[9], Hamdi Mbarek[78], Rachel L. Maguire[21,79], Terence Dwyer[80], Martine Vrijheid[11,53,54], Luigi Bouchard[81,82], Andrea A. Baccarelli[83,84], Lisa A. Croen[85], Wilfried Karmaus[57], Denise Anderson[32], Maaike de Vries[33], Sylvain Sebert[61,62,86], Juha Kere[87,88,89], Robert Karlsson[36], Syed Hasan Arshad[27,90], Esa Hämäläinen[91], Michael N. Routledge[92], Dorret I. Boomsma[39,93], Andrew P. Feinberg[94], Craig J. Newschaffer[95], Eva Govarts[40], Matthieu Moisse[96,97], M. Daniele Fallin[98], Erik Melén[35,99], Andrew M. Prentice[67], Eero Kajantie[100,101,102], Catarina Almqvist[36,103], Emily Oken[104], Dana Dabelea[105], H. Marike Boezen[33], Phillip E. Melton[106,107], Rosalind J. Wright[25], Gerard H. Koppelman[30], Letizia Trevisi[108], Marie-France Hivert[55,104,109], Jordi Sunyer[11,53,54,76], Monica C. Munthe-Kaas[110,111], Susan K. Murphy[112], Eva Corpeleijn[3], Joseph Wiemels[113], Nina Holland[9], Zdenko Herceg[12], Elisabeth B. Binder[16,114], George Davey Smith[1,2], Vincent W.V. Jaddoe[5,6,7], Rolv T. Lie[13,115], Wenche Nystad[116], Stephanie J. London[15], Debbie A. Lawlor[1,2], Caroline L. Relton[1,2], Harold Snieder[3] & Janine F. Felix[5,6,7]

[1]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [2]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [3]University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands. [4]Division of Human Nutrition and Health, Wageningen University, Wageningen, The Netherlands. [5]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [6]Department of Epidemiology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [7]Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [8]School of Oral and Dental Sciences, University of Bristol, Bristol, UK. [9]Children's Environmental Health Laboratory, Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, USA. [10]Department of Epidemiology, Geisel School of Medicine at Dartmouth College, Hanover, NH, USA. [11]ISGlobal, Bacelona Institute for Global Health, Barcelona, Spain. [12]Epigenetics Group, International Agency for Research on Cancer, Lyon, France. [13]Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo, Norway. [14]Oslo Centre for Biostatisitcs and Epidemology, Oslo University Hospital, Oslo, Norway. [15]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Service, Research Triangle Park, Durham, NC, USA. [16]Department of Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany. [17]Department of Epidemiology, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [18]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [19]HDF Comprehensive Cancer Center, University of California, San Francisco, CA, USA. [20]Computational Biology and Informatics, UCSF, San Francisco, CA, USA. [21]Department of Biological Sciences, North Carolina State University, Raleigh, NC, USA. [22]Center for Human Health and the Environment, North Carolina State University, Raleigh, NC, USA. [23]Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089, USA. [24]Centre de recherche du Centre hospitalier universitaire de Sherbrooke, Sherbrooke, QC, Canada. [25]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [26]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [27]Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. [28]Human Development and Health, Faculty of Medicine, University of Southampton, Southampton General Hospital, Southampton, UK. [29]Department of Environmental Health, Rollins School of Public Health at Emory University, Atlanta, GA, USA. [30]University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, Groningen Research Institute for Asthma and COPD, Groningen, The Netherlands. [31]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands. [32]Telethon Kids Institute, University of Western Australia, Perth, Australia. [33]University of Groningen, University Medical Center Groningen, Department of Epidemiology and Groningen Research Institute for Asthma and

COPD (GRIAC), Groningen, The Netherlands. [34]Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment & Health, School of Public Health, Imperial College London, London, UK. [35]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [36]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [37]Helsinki Collegium for Advanced Studies, University of Helsinki, Helsinki, Finland. [38]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [39]Department of Biological Psychology, Netherlands Twin Register, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [40]VITO - Health, Mol, Belgium. [41]Theoretical Physics, Faculty of Sciences, Hasselt University, Hasselt, Belgium. [42]Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium. [43]Research Unit for Gynaecology and Obstetrics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark. [44]Department of Pediatrics, Division of Respiratory Medicine and Allergology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [45]Department of Pediatrics, Division of Neonatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands. [46]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, Durham, NC, USA. [47]Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [48]Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium. [49]MRC/PHE Centre for Environment and Health School of Public Health Imperial College London, St Mary's Campus, Norfolk Place, London, UK. [50]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [51]Bioinformatics Research Center, North Carolina State University, Raleigh, NC, USA. [52]Bioinformatics and Genomics Program, Centre for Genomic Regulation (CRG), Barcelona, Spain. [53]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [54]CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain. [55]Department of Medicine, Universite de Sherbrooke, Sherbrooke, QC, Canada. [56]Department of Public Health Sciences, School of Medicine, University of California Davis MIND Institute, Sacramento, CA, USA. [57]Division of Epidemiology, Biostatistics and Environmental Health, School of Public Health, University of Memphis, Memphis, TN, USA. [58]Children's Environmental Health & Disease Prevention Research Center at Dartmouth, Hanover, NH, USA. [59]Institute for Risk Assessment Sciences, Utrecht University, Utrecht, The Netherlands. [60]Medical School, University of Western Australia, Perth, Australia. [61]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, 90014 Oulu, Finland. [62]Biocenter Oulu, University of Oulu, Oulu, Finland. [63]Unit of Primary Care, Oulu University Hospital, Oulu, Finland. [64]Center for Occupational and Environmental Medicine, Stockholm County Council, Stockholm, Sweden. [65]College of Veterinary Medicine, Michigan State University, East Lansing, MI, USA. [66]Obstetrics and Gynaecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland. [67]Medical Research Council Unit The Gambia at the London School of Hygiene and Tropical Medicine, London, UK. [68]Department of Women and Children's Health, King's College London, London, UK. [69]Novo Nordisk Foundation Center for Basic Metabolic Research, Section of Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [70]Department of Public Health, Section of Epidemiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [71]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [72]Department of Public Health & Primary Care, Leuven University, Leuven, Belgium. [73]School of Food Sciences and Nutrition, University of Leeds, Leeds, UK. [74]Department of Medical Sciences, University of Turin, Turin, Italy. [75]AOU Citta della Salute e della Sceinza, CPO Piemonte, Turin, Italy. [76]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain. [77]Center for Environmental Research and Children's Health, School of Public Health, University of California, Berkeley, CA, USA. [78]Department of Biological Psychology, Amsterdam Public Health Research Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [79]Department of Community and Family Medicine, Duke University Medical Center, Raleigh, NC, USA. [80]The George Institute for Global Health, Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK. [81]Department of Biochemistry, Université de Sherbrooke, Sherbrooke, QC, Canada. [82]ECOGENE-21 Biocluster, Chicoutimi Hospital, Saguenay, QC, Canada. [83]Laboratory of Precision Environmental Biosciences, Columbia University Mailman School of Public Health, New York, NY, USA. [84]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University, New York, NY, USA. [85]Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. [86]Department for Genomics of Common Diseases, School of Public Health, Imperial College London, London, UK. [87]Folkhälsan Institute of Genetics, Helsinki, and Research Programs Unit, Molecular Neurology, University of Helsinki, Helsinki, Finland. [88]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden. [89]School of Basic and Medical Biosciences, King's College London, Guy's Hospital, London, UK. [90]David Hide Asthma and Allergy Research Centre, Isle of Wight, UK. [91]HUSLAB and the Department of Clinical Chemistry, University of Helsinki, Helsinki, Finland. [92]LICAMM, School of Medicine, Univeristy of Leeds, Leeds, UK. [93]Amsterdam Public Health Institute, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [94]Center for Epigenetics, Johns Hopkins University School of Medicine, Baltimore, MA, USA. [95]AJ Drexel Autism Institute, Drexel University, Philadelphia, PA, USA. [96]KU Leuven – University of Leuven, Department of Neurosciences, Experimental Neurology and Leuven Institute for Neuroscience and Disease (LIND), Leuven, Belgium. [97]VIB, Center for Brain & Disease Research, Laboratory of Neurobiology, Leuven, Belgium. [98]Department of Mental Health, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA. [99]Sachs' Children's Hospital, Stockholm, Sweden. [100]National Institute for Health and Welfare, Helsinki and Oulu, Oulu, Finland. [101]Hospital for Children and Adolescents, Helsinki University Hospital and University of Helsinki, Helsinki, Finland. [102]PEDEGO Research Unit, MRC Oulu, Oulu University Hospital and University of Oulu, Oulu, Finland. [103]Pediatric Allergy and Pulmonology Unit at Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden. [104]Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute, Boston, MA, USA. [105]Department of Epidemiology, Colorado School of Public Health, and Department of Pediatrics, University of Colorado School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [106]Centre for Genetic Origins of Health and Disease, School of Biomedical Sciences, University of Western Australia, Perth, Australia. [107]School of Pharmacy and Biomedical Sciences, Curtin University, Perth, Australia. [108]Department of Global Health and Social Medicine, Harvard Medical School, Boston, MA, USA. [109]Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [110]Norwegian Institute of Public Health, Oslo, Norway. [111]Department of Pediatric Oncology and Hematology, Oslo University Hospital, Oslo, Norway. [112]Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC, USA. [113]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA. [114]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Altanta, GA, USA. [115]Department of Global Public Health and Primary Care, University of Bergen, Bergen, Norway. [116]Department for Non-Communicable Diseases, Norwegian Institute for Public Health, Oslo, Norway. These authors jointly supervised this work: Debbie A. Lawlor, Caroline L. Relton, Harold Snieder, Janine F. Felix.

Merid et al. Genome Medicine (2020) 12:25
https://doi.org/10.1186/s13073-020-0716-9

Genome Medicine

RESEARCH Open Access

Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age

Simon Kebede Merid[1,2†], Alexei Novoloaca[3†], Gemma C. Sharp[4,5†], Leanne K. Küpers[5,6,7†], Alvin T. Kho[8†], Ritu Roy[9,10], Lu Gao[11], Isabella Annesi-Maesano[12], Pooja Jain[13,14], Michelle Plusquin[13,15], Manolis Kogevinas[16,17,18,19], Catherine Allard[20], Florianne O. Vehmeijer[21,22], Nabila Kazmi[4,5], Lucas A. Salas[23], Faisal I. Rezwan[24], Hongmei Zhang[25], Sylvain Sebert[26,27,28], Darina Czamara[29], Sheryl L. Rifas-Shiman[30], Phillip E. Melton[31,32], Debbie A. Lawlor[4,5,33], Göran Pershagen[1,34], Carrie V. Breton[11], Karen Huen[35], Nour Baiz[12], Luigi Gagliardi[36], Tim S. Nawrot[13,37], Eva Corpeleijn[7], Patrice Perron[20,38], Liesbeth Duijts[21,22], Ellen Aagaard Nohr[39], Mariona Bustamante[16,17,18], Susan L. Ewart[40], Wilfried Karmaus[25], Shanshan Zhao[41], Christian M. Page[42], Zdenko Herceg[3], Marjo-Riitta Jarvelin[26,27,43,44], Jari Lahti[45,46], Andrea A. Baccarelli[47], Denise Anderson[48], Priyadarshini Kachroo[49], Caroline L. Relton[4,5,33], Anna Bergström[1,34], Brenda Eskenazi[50], Munawar Hussain Soomro[12], Paolo Vineis[51], Harold Snieder[7], Luigi Bouchard[20,52,53], Vincent W. Jaddoe[21,22], Thorkild I. A. Sørensen[4,54,55], Martine Vrijheid[16,17,18], S. Hasan Arshad[56,57], John W. Holloway[58], Siri E. Håberg[42], Per Magnus[42], Terence Dwyer[59,60], Elisabeth B. Binder[29,61], Dawn L. DeMeo[49], Judith M. Vonk[7,62], John Newnham[63], Kelan G. Tantisira[49], Inger Kull[2,64], Joseph L. Wiemels[65], Barbara Heude[66], Jordi Sunyer[16,17,18,19], Wenche Nystad[42], Monica C. Munthe-Kaas[42,67], Katri Räikkönen[42], Emily Oken[30], Rae-Chi Huang[48], Scott T. Weiss[49], Josep Maria Antó[16,17,18,19], Jean Bousquet[68,69], Ashish Kumar[1,70,71], Cilla Söderhäll[72], Catarina Almqvist[73,74], Andres Cardenas[75], Olena Gruzieva[1,34], Cheng-Jian Xu[76], Sarah E. Reese[41], Juha Kere[77,78], Petter Brodin[72,79,80], Olivia Solomon[35], Matthias Wielscher[43], Nina Holland[35], Akram Ghantous[3], Marie-France Hivert[20,30,81], Janine F. Felix[21,22], Gerard H. Koppelman[76], Stephanie J. London[41†] and Erik Melén[1,2,82*†]

* Correspondence: erik.melen@ki.se
†Simon Kebede Merid, Alexei Novoloaca, Gemma C. Sharp, Leanne K. Küpers and Alvin T. Kho are shared first authors.
†Erik Melén and Stephanie J. London are shared senior authors.
[1]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[2]Department of Clinical Sciences and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden
Full list of author information is available at the end of the article

Abstract

Background: Preterm birth and shorter duration of pregnancy are associated with increased morbidity in neonatal and later life. As the epigenome is known to have an important role during fetal development, we investigated associations between gestational age and blood DNA methylation in children.

Methods: We performed meta-analysis of Illumina's HumanMethylation450-array associations between gestational age and cord blood DNA methylation in 3648 newborns from 17 cohorts without common pregnancy complications, induced delivery or caesarean section. We also explored associations of gestational age with DNA methylation measured at 4–18 years in additional pediatric cohorts. Follow-up analyses of DNA methylation and gene expression correlations were performed in cord blood. DNA methylation profiles were also explored in tissues relevant for gestational age health effects: fetal brain and lung.

Results: We identified 8899 CpGs in cord blood that were associated with gestational age (range 27–42 weeks), at Bonferroni significance, $P < 1.06 \times 10^{-7}$, of which 3343 were novel. These were annotated to 4966 genes. After restricting findings to at least three significant adjacent CpGs, we identified 1276 CpGs annotated to 325 genes. Results were generally consistent when analyses were restricted to term births. Cord blood findings tended not to persist into childhood and adolescence. Pathway analyses identified enrichment for biological processes critical to embryonic development. Follow-up of identified genes showed correlations between gestational age and DNA methylation levels in fetal brain and lung tissue, as well as correlation with expression levels.

Conclusions: We identified numerous CpGs differentially methylated in relation to gestational age at birth that appear to reflect fetal developmental processes across tissues. These findings may contribute to understanding mechanisms linking gestational age to health effects.

Keywords: Development, Epigenetics, Gestational age, Preterm birth, Transcriptomics

Background

Preterm birth (birth before 37 weeks' gestation) is associated with increased neonatal morbidity and mortality [1, 2], as well as later health [3–6]. In children born at very young gestational ages, bronchopulmonary dysplasia, retinopathy and neurodevelopmental impairment are major health challenges [7–12]. Lower lung function is observed in children born moderately preterm, i.e. between 32 and 36 completed weeks, compared to those born at term [13]. Even variation in gestational age within the normal range (37–41 weeks) is related to various health outcomes, including neurological and cognitive development [14–17] and respiratory disease [4]. Mechanisms for many of these findings are not well understood.

The epigenome is known to have an important role during fetal development. The best studied epigenetic modification is methylation. DNA methylation patterns have been associated with environmental factors relevant to preterm birth, including smoking, air pollution exposure, microbial and maternal nutritional factors [18–22]. Such exposure-related epigenetic patterns potentially influence gene expression profiles and/or susceptibility to chronic disease during the life-course [23, 24]. Further, DNA methylation in whole blood at birth may also reflect development across fetal life. It is possible that DNA methylation changes at birth may contribute to the myriad immediate and late health outcomes that have been associated with gestational age.

Knowledge about DNA methylation and gene expression profiles associated with length of gestation may help to better understand both the molecular basis of abnormal processes related to prematurity as well as normal human development. Several studies have reported associations of gestational age among both term and preterm births with cord blood DNA methylation [25–29]. In the largest EWAS to date ($n = 1753$ newborns), 5474 CpGs in cord blood were associated with gestational age [30]. While these individual studies have identified widespread associations of DNA methylation patterns at birth with gestational age, meta-analysis of results from multiple individual cohorts increases sample size and, thus, greatly increases power to detect robust differential methylation signals.

We examined DNA methylation levels in newborns in relation to gestational age in a large-scale meta-analysis and also examined functional effects on expression of nearby genes of potential relevance for later health. We meta-analysed harmonized cohort specific EWAS results of the association of gestational age with cord blood DNA methylation levels from the Pregnancy And Childhood Epigenetics (PACE) Consortium of pregnancy and childhood cohorts [31]. We also examined associations with continuous gestational age limited to term newborns. CpGs that were differentially methylated in cord blood in relation to gestational age were then analysed

in two fetal tissues (lung and brain), with relevance for health impacts of low gestational age [7–12]. We conducted analyses to explore whether associations of CpG methylation with gestational age persisted in older children aged 4–18 years. DNA methylation status at the identified CpGs was analysed for association with gene expression patterns of nearby genes in cord blood during different developmental stages. Finally, we performed pathway and functional network analysis of identified genes to gain insight into the biological implications of our findings.

Methods

Figure 1 gives an outline of the design of this study.

Study population

A total of 11,000 participants in 26 independent cohorts were included in our study. In the "all births model" meta-analysis, we included $n = 6885$ newborns from 20 cohorts. In our main "no complications model", we excluded participants with maternal complications (maternal pre-eclampsia or diabetes or hypertension) and caesarean section delivery or delivery start with induction, leaving 3648 newborns from 17 cohorts for this analysis (Additional file 1: Table S1). For the additional look-up of persistent differential methylation at later ages, we used participants from 4 cohorts with whole blood DNA methylation in early childhood (4–5 years; $n = 453$), 5 cohorts with whole blood DNA methylation at school age (7–9 years; $n = 899$) and 5 cohorts with whole blood DNA methylation in adolescence (16–18 years; $n = 1129$). Detailed methods for each cohort are provided in Additional file 2: Supplementary information. All cohorts acquired ethics approval and informed consent from participants prior to data collection through local ethics committees (Additional file 2: Supplementary information).

Gestational age

In each cohort, information on gestational age at birth was obtained from birth certificates ($n = 725$), medical records using ultrasound estimation ($n = 1931$), or last menstrual period date ($n = 468$), or combined estimate from ultrasound and last menstrual period date ($n = 6630$), or otherwise from self-administrated questionnaires ($n = 1246$). Gestational age was analysed in days. Women with a gestational age of more than 42 weeks (294 days) were excluded from all models. Additionally, multiple births were also excluded from the analysis.

Methylation measurements and quality control

DNA methylation from newborns and older children was measured using the Illumina450K platform. Each cohort conducted their own quality control and normalization of DNA methylation data, as detailed in

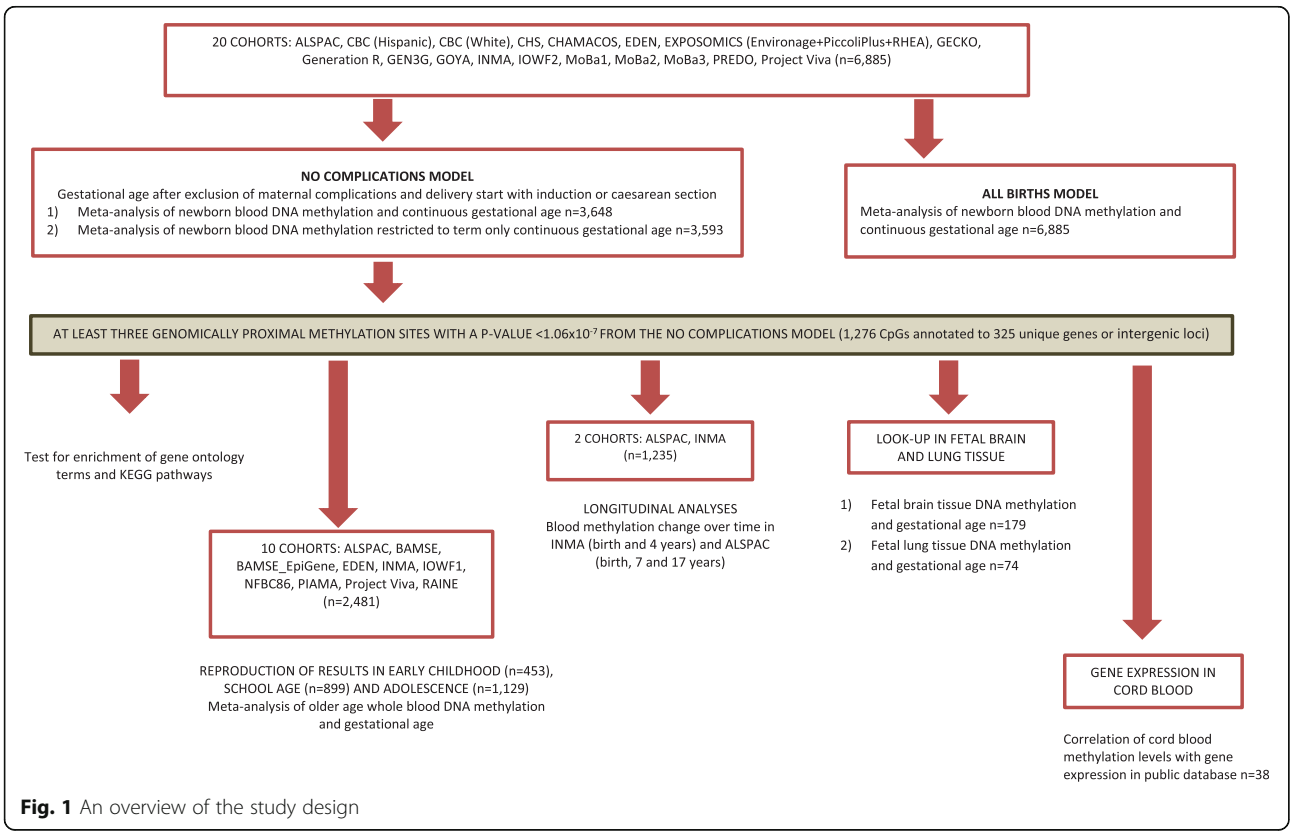


Fig. 1 An overview of the study design

Additional file 1: Table S2. Cohorts corrected for batch effects in their data using surrogate variables, ComBat [32], or by including a batch covariate in their models. To reduce the impact of severe outliers in the DNA methylation data on the meta-analysis, cohorts trimmed the methylation beta values by removing, for each CpG, observations more than three times the interquartile range below the 25th percentile or above the 75th percentile [33]. Cohorts retained all CpGs that passed quality control and removed CpGs that were mapped to the X ($n = 11{,}232$) or Y ($n = 416$) chromosomes and control probes ($n = 65$), leaving a maximum total of 473,864 CpGs included in the meta-analysis.

Cohort-specific statistical analyses

Each cohort performed independent EWAS according to a common, pre-specified analysis plan. Robust linear regression (rlm in the MASS R package [34]) was used to model gestational age as the exposure and DNA methylation beta values as the outcome. In the primary analysis, gestational age was used as a continuous variable excluding cohorts that had term-only infants. In secondary models, we modeled term-only children defined as a gestational age ≥ 37 weeks (≥ 259 days), but less or equal with 42 weeks. All models were adjusted for sex, maternal age (years), maternal social class (variable defined by each individual cohort; Additional file 1: Table S2), maternal smoking status (the preferred categorization was into three groups: no smoking in pregnancy, stopped smoking in early pregnancy, smoking throughout pregnancy, but a binary categorization of any versus no smoking was also acceptable), parity (the preferred categorization was into two groups: no previous children, one or more previous children), birth weight in grams, age of the child (years) included for older children, batch or surrogate variables. Optionally, cohorts could include ancestry, and/or selection covariates, if relevant to their study. We also adjusted for potential confounding by cell type using estimated cell type proportions calculated from a cord blood cell type reference panel [35] for newborn cohorts or the adult blood cell type reference panel [36] for cohorts with older children using the *estimateCellCounts* function in the *minfi* R package [37].

Meta-analysis

We performed fixed-effects meta-analysis weighted by the inverse of the variance with METAL [38]. A shadow meta-analysis was also conducted independently by a second study group (see author contribution) and the results were compared [39] (and confirmed). All downstream analyses were conducted using R version 2.5.1 or later [40]. Multiple testing was accounted for by applying the Bonferroni correction level for 473,864 tests ($P < 1.06 \times 10^{-7}$). A random effects model was performed using the METASOFT tool [41]. We explored heterogeneity between studies using the I^2 statistic [42]. A priori, we defined $I^2 > 50\%$ as reflecting a high level of between-study variation. In case of $I^2 > 50\%$, we replaced values with random effects estimates as these are attenuated in the face of heterogeneity and thus more conservative. To focus functional analyses and bioinformatics efforts on genes and loci that were found to be robustly associated with gestational age, we selected regions that had at least three adjacent Bonferroni significant CpGs ($P < 1.06 \times 10^{-7}$) [43]. Genome-wide DNA methylation meta-analysis summary statistics corresponding to the main analysis presented in this manuscript are available at figshare (https://doi.org/10.6084/m9.figshare.11688762.v1) [44].

Analyses of differentially methylated regions

Differentially methylated regions (DMRs) were identified using two methods available for meta-analysis results comb-p [45] and DMRcate [46]. Input parameters used for the DMR calling in both algorithms are provided in Additional file 2: Supplementary information. Comb-p uses a one-step Šidák correction [45] and DMRcate uses an FDR correction [46] per default. The selected regions were defined based on the following criteria: the minimum number of CpGs in a region had to be 2, regional information can be combined from probes within 1000 bp and the multiple-testing corrected $P < 0.01$ (Šidák-corrected $P < 0.01$ from comb-p and FDR < 0.01 from DMRcate).

Analyses of embryonic DNA methylation

DNA methylation from lung tissue of 74 foetuses (estimated ages 59 to 122 days post conception [47]) were used for analyses of differentially methylated CpGs (three or more adjacent Bonferroni significant CpGs, $P < 1.06 \times 10^{-7}$; $n = 1276$) from the newborn meta-analysis. A linear regression model adjusted for sex and in utero smoke exposure (IUS) was applied. A Bonferroni look-up level correction (0.05/1030; $P < 4.85 \times 10^{-5}$) considered as significance threshold, followed by a comparison of the direction of effect with that in the cord blood meta-analysis. We also performed look-up analyses of selected 1276 CpGs in another organ, fetal brain tissue, from 179 foetuses collected between 23 and 184 days post-conception [48]. For these analyses, we kept the available Bonferroni correction $P < 1.06 \times 10^{-7}$ as significance threshold, followed by a comparison of the direction of effect with that in the cord blood meta-analysis.

Look-up analyses in older ages

Differentially methylated CpGs (three or more adjacent CpGs below the Bonferroni correction $P < 1.06 \times 10^{-7}$; $n = 1276$) from the newborn meta-analyses were

analysed with a look-up approach using data from four early childhood, five school age, and five adolescence cohorts. Cohorts included the same covariates in these analyses as in the cord blood analyses and child age. We performed fixed effects inverse variance weighted meta-analyses using METAL [38] for these three age groups. For this hypothesis-driven analysis, CpG methylation association with gestational age was considered statistically significant at nominal $P < 0.05$, followed by a comparison of the direction of effect with that in the cord blood meta-analysis.

Longitudinal analysis

Longitudinal DNA methylation data from birth to early childhood and from birth to adolescence were analysed for the three or more adjacent Bonferroni significant 1276 CpGs found to be associated with gestational age. DNA methylation from two time points (birth and 4 years) in INMA and three time points (birth, 7 and 17 years) in ALSPAC were analysed separately. To estimate changes in DNA methylation, we applied linear mixed models with repeated measurement taking into account the within-person time effect. The models were adjusted for covariates and estimated cell count similar to cross-sectional analysis. Interaction terms between age and gestational age were included in the model to capture differences in methylation change between birth and 4 years, birth and 7 years and 7 and 17 years per day increase in gestational age at delivery, respectively. The stable CpGs that did not change significantly from birth to adolescence had no association with age (at nominal $P < 0.05$), and no interaction between gestational age and childhood age (at nominal $P < 0.05$).

Enrichment and functional analysis

CpGs were annotated using *FDb.InfiniumMethylation.hg19* R package, with enhanced annotation for nearest genes within 10 Mb of each site, as previously described [20]. Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses were performed using the overrepresentation analysis (ORA) tool ConsensusPathDB (http://consensuspathdb.org/ [49, 50]). *P* values for enrichment were adjusted for multiple testing using the FDR method.

DNA methylation in relation to gene expression

Correlations between DNA methylation and gene expression levels were tested using paired DNA methylation and gene expression data in publicly available datasets. We tested transcript levels of genes within a 500-kb region of the 1276 three adjacent CpGs (250 kb upstream and 250 kb downstream). The mRNA gene expression (Affymetrix Human Transcriptome Array 2.0) and methylation (Illumina Infinium® HumanMethylation450 BeadChip assay) were measured in cord-blood samples from 38 newborns [51–53]. First, we created residuals for mRNA expression and residuals for DNA methylation and used linear regression models to evaluate correlations between expression residuals and DNA methylation residuals. These residual models were adjusted for covariates, estimated white blood cell proportions, and technical variation. We corrected these analyses for multiple testing using Bonferroni correction.

Results

Study characteristics

We meta-analysed Illumina's HumanMethylation450-array results from 17 independent cohorts with data on newborn DNA methylation status, and 10 cohorts with data on DNA methylation in older children (age 4 to 18 years), including 4 cohorts with DNA methylation data both at birth and at an older age (Fig. 1). Table 1 summarizes the characteristics of participating cohorts. A summary of methods used by each cohort is provided in Additional file 1: Tables S1 and S2. In our main "no complications" model, we excluded participants exposed to maternal pregnancy complications (maternal diabetes, hypertension or pre-eclampsia) and whose labour was induced or who were delivered by caesarean section. With continuous gestational age in the number of days as the exposure (gestational age range 186–294 days corresponding to 27–42 weeks), we analysed results from 3648 newborns and from 2481 older children. This model was selected as the main model because associations of DNA methylation with gestational age related to pregnancy complications or potentially influenced by obstetric interventions may be less reflective of normal developmental processes than newborns with spontaneous uncomplicated delivery. However, we also analysed a larger dataset of 6885 newborns from 20 independent cohorts, including pregnancies with pregnancy complications and obstetric interventions, referred to as the "all births model" (see below).

Associations between gestational age and newborn DNA methylation

We identified 8899 CpGs in cord blood that were associated with gestational age (range 27–42 weeks), at Bonferroni significance, $P < 1.06 \times 10{-}7$, of which 3343 were novel. These were annotated to 4966 genes. CpGs associated with gestational age had a modest predominance of negative (60%) versus positive (40%) direction of effect, with an overall absolute median difference in mean methylation of 0.36% per gestational week, IQR = [0.26%–0.49%] (Fig. 2a). In general, results were highly homogeneous; evidence of high between-study heterogeneity, using a criterion of $I^2 > 50\%$, was seen for only

Table 1 Characteristics of each cohort included in the association meta-analysis between gestational age (GA) and DNA methylation in newborns and older children

Study population	Cohort	*N*	*N*, preterm*	*N*, term	Age mean (SD)	Maternal age mean (SD)	Mean GA (days)	SD GA	Min GA	Max GA	Ethnicity
Newborn	ALSPAC** [29]	249	10	239	0	29.8 (4.6)	277	10.78	224	294	European
	CBC (Hispanic) [54]	128	10	118	0	27.3 (5.8)	273	17.70	196	294	Hispanic
	CBC (European) [54]	132	11	121	0	31.9 (5.7)	273	16.10	189	294	European
	CHS [55]	120	7	113	0	29.4 (5.6)	277	11.20	230	294	Mixed
	CHAMACOS [56]	110	11	99	0	25.3 (5.0)	272	10.66	210	294	Hispanic
	EDEN [57]	100	2	98	0	30.8 (5.0)	276	10.11	217	287	European
	EXPOSOMICS (Environage + PiccoliPlus + RHEA) [58]	252	17	235	0	30.5 (4.8)	273	10.50	217	294	European
	Generation R [59]	486	22	464	0	31.9 (4.2)	280	9.00	239	294	European
	INMA [60]	134	2	132	0	30.5 (4.1)	278	9.57	234	286	European
	IOW F2 [61]	93	2	91	0	23.2 (2.6)	278	10.95	236	294	European
	MoBa1** [30]	749	18	731	0	29.9 (4.3)	279	10.36	209	294	European
	MoBa2** [30]	460	15	445	0	30.0 (4.5)	278	10.49	209	294	European
	MoBa3 [20]	177	3	174	0	29.6 (4.4)	279	10.38	199	294	European
	PREDO [62]	308	5	303	0	33.4 (5.7)	278	11.20	186	294	European
	Project Viva [63]	150	3	147	0	33.2 (4.5)	278	10.11	216	294	European
	Meta-analysis	3648	138								
Early childhood	BAMSE [64]	145	10	135	4.3 (0.2)	31.2 (4.4)	275	16.22	187	293	European
	EDEN [64]	89	2	87	5.6 (0.1)	30.8 (5.1)	276	9.23	245	287	European
	INMA [64]	71	1	70	4.4 (0.2)	30.6 (4.3)	279	8.70	249	288	European
	PIAMA [64]	148	4	144	4.1 (0.2)	30.6 (3.6)	278	10.51	233	294	European
	Meta-analysis	453	17								
School age	ALSPAC [29]	273	12	261	7.5 (0.1)	29.9 (4.6)	277	10.99	224	294	European
	BAMSE [64]	141	10	131	8.4 (0.4)	31.4 (4.5)	276	15.96	197	293	European
	BAMSE_EpiGene [64]	232	8	224	8.3 (0.5)	30.8 (4.4)	278	11.47	209	294	European
	PIAMA [64]	134	3	131	8.1 (0.3)	30.5 (3.6)	278	10.61	233	294	European
	Project Viva [63]	119	2	117	7.8 (0.7)	33.5 (4.4)	278	10.32	216	294	European
	Meta-analysis	899	35								
Adolescence	ALSPAC [29]	272	13	259	17.2 (1.0)	29.9 (4.6)	277	11.04	224	294	European
	BAMSE [64]	159	7	152	16.7 (0.4)	31.2 (4.4)	278	12.70	187	294	European
	IOW F1 [61]	97	2	95	17.1 (0.5)	27.1 (5.1)	280	9.83	238	294	European
	NFBC86 [65]	287	9	276	16.1 (0.4)	29.0 (5.1)	280	8.65	237	294	European
	RAINE [66]	314	9	305	17.0 (0.3)	29.0 (5.8)	274	11.90	196	294	European
	Meta-analysis	1129	40								

*Preterm birth categorized as GA less than 37 full weeks or 259 days and as term greater than 37 weeks or 259 days (but less than 42 full weeks). **This study was included previous EWAS of gestational age [29, 30]. Cohort details and references can be found at Additional file 2 and in Felix et al. [31]

319 of the 8899 CpGs (Additional file 1: Table S3). Leave one out analyses did not indicate an influential effect on meta-analysis results of any single study. However, we replaced fixed effects values with random effects estimates for those CpGs with between study $I^2 > 50\%$, as these are more conservative in the case of heterogeneity.

Differentially methylated CpGs spanned all chromosomes (Fig. 2b). The CpG with the lowest *P* value ($P = 2.7 \times 10^{-129}$ for cg16103712; Table 2) was annotated to *MATN2* on chr 8, and the difference in mean methylation at this CpG was 2.13% lower per additional gestational week (equal to 0.30% per day). The CpG with the largest negative association was cg04347477, annotated to *NCOR2* on chr 12 (Table 3), with a lower mean methylation of 2.53% per additional gestational week. *B3GALT4* (chr 6) had the largest number of significant CpGs negatively associated with gestational age (21 out of 52

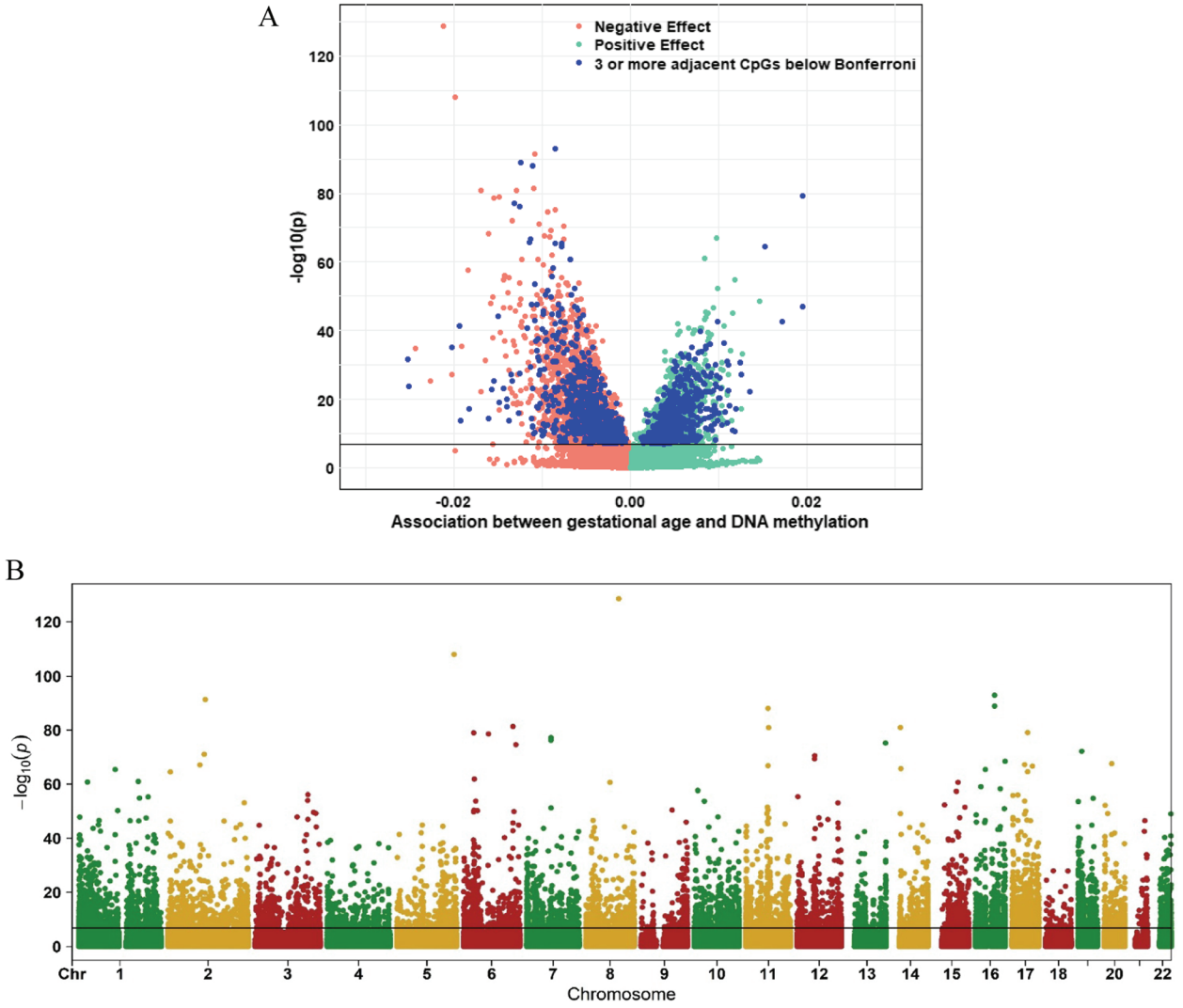


Fig. 2 **A**, **B** Volcano (**A**) and Manhattan (**B**) plots for the meta-analysis of gestational age and offspring DNA methylation association at birth, after adjustment for covariates and estimated cell proportions. The effect size represents methylation change per gestational week

(40%) tested CpGs annotated to *B3GALT4*). The largest positive association was observed for cg13036381 annotated to *LOC401097* (chr 3) (Table 3) with a difference in mean methylation of 1.95% per additional gestational week. *DDR1* (chr 6) had the largest number of significant CpGs positively associated with gestational age (26/95 (27%) CpGs). A complete list of associated CpGs is presented in Additional file 1: Table S3 and the CpG variation across cohorts in Additional file 3: Figure S1 (top CpGs).

We performed a sensitivity analysis by excluding cohorts that were included in previous EWAS of gestational age [29, 30] (three cohorts: MoBa1, MoBa2 and ALSPAC) in order to evaluate associations not driven by previous results, and found a high correlation ($r = 0.89$) of effect estimates (Additional file 3: Figure S2) compared with results from all cohorts included in the no complication model.

Next, we performed a meta-analysis of the larger dataset of 6885 participants from 20 studies without excluding maternal complications and caesarean section delivery or induced delivery. In this "all births model", 17,095 CpGs located in or near 7931 genes were associated with gestational age after Bonferroni correction ($P < 1.06 \times 10^{-7}$). Not surprisingly given the higher levels of statistical significance in this much larger data set, we found somewhat more between-study heterogeneity than in the no complications model, but high levels ($I^2 > 50\%$) were observed for only 1784 out of these 17,095 CpGs (Additional file 1: Table S4). We also observed a considerable overlap of CpGs between the two models with 93% of the 8899 CpGs in the no complication model also

Table 2 The top 10 Bonferroni-significant CpGs from the meta-analysis on the association between continuous GA and offspring DNA methylation at birth adjusted for estimated cell proportions

CpGID	Chr	Genomic coordinates	Gene (Illumina annotation)	Relation to island	Distance to nearest gene	UCSC known gene	Coefficient*	*P* value	Direction of effect in each cohort**
cg16103712	8	99,023,869	*MATN2*	OpenSea	7355	*MATN2*	− 0.0030	2.70E−129	----------
cg04685228	5	172,462,626		OpenSea	726	*ATP6V0E1*	− 0.0028	8.55E−109	------?--------
cg04276536	16	57,567,813	*CCDC102A*	N_Shelf	0	*CCDC102A*	− 0.0012	1.20E−93	------?--------
cg19744173	2	112,913,178	*FBLN7*	N_Shelf	0	*FBLN7*	− 0.0016	4.91E−92	----------
cg27518892	16	57,566,936	*CCDC102A*	N_Shelf	0	*CCDC102A*	− 0.0018	1.29E−89	----------
cg13924996	11	67,053,829	*ADRBK1*	S_Shore	0	*ADRBK1*	− 0.0016	8.59E−89	------?--------
cg04494800	6	149,775,853	*ZC3H12D*	N_Shore	1923	*ZC3H12D*	− 0.0016	4.52E−82	------?--------
cg27295118	14	22,902,226		OpenSea	− 500	*AK125397*	− 0.0024	1.20E−81	------?--------
cg26433582	11	68,848,232	*TPCN2*	N_Shore	917	*TPCN2*	− 0.0019	1.31E−81	------?--------
cg18183624	17	47,076,904	*IGF2BP1*	S_Shore	0	*IGF2BP1*	0.0028	8.36E−80	++++++++++++++

*Coefficient corresponding to methylation change per additional day of gestational age
**Order of included cohorts in the meta-analysis: MoBa1, MoBa2, MoBa3, EDEN, EXPOSOMICS (Environage+PiccoliPlus+RHEA), CHS, IOWF2, Generation R, Project Viva, CBC (Hispanic), CBC (White), ALSPAC, PREDO, CHAMACOS and INMA."?" Means that CpG was not measured in that cohort

reaching Bonferroni significance in the all birth model and showing the same direction of effect.

CpG localization and regulatory region analyses

The 8899 differentially methylated CpGs in relation to continuous gestational age in the no complications model were enriched for localization to CpG island shores (33% of the 8899 CpGs are in shores, whereas 23% of all CpGs on the 450 K array are in shores, $P_{\text{enrichment}} = 4.1\times 10^{-100}$, Fig. 3), open sea (45% versus 37%, $P_{\text{enrichment}} = 1.4 \times 10^{-63}$), enhancers (37% versus 22%, $P_{\text{enrichment}} = 1.05 \times 10^{-236}$), DNase hypersensitivity sites (18% versus 12%, $P_{\text{enrichment}} = 1.3\times 10^{-56}$) and CpG island shelves (12% versus 10%, $P_{\text{enrichment}} = 1.2 \times 10^{-11}$) (Fig. 3). In contrast, we found relative depletion in CpG islands (10% versus 31%, $P_{\text{enrichment}} = 2.2 \times 10^{-308}$), FANTOM 4 promoters (2.3% versus 6.7%, $P_{\text{enrichment}} = 6.7 \times 10^{-79}$) and promoter-associated regions (11% versus 19%, $P_{\text{enrichment}} = 2.2 \times 10^{-104}$).

Analysis restricted to term-births

To evaluate whether observed DNA methylation differences in relation to continuous gestational age were driven by preterm birth, we repeated the no complication model including only infants born at term (gestational age 37 to 42 weeks). In this analysis, we meta-analysed results from 18 cohorts (one additional cohort with term-birth data only was

Table 3 The top 10 Bonferroni-significant CpGs ranked by the magnitude of positive and negative effect (5 CpGs each) from the meta-analysis on the association between continuous GA and offspring DNA methylation at birth adjusted for estimated cell proportions

CpGID	Chr	Genomic coordinates	Gene (Illumina annotation)	Relation to island	Distance to nearest gene	UCSC known gene	Coefficient*	*P* value	Direction of effect in each cohort**
cg13036381	3	1.6E+ 08	*LOC401097*	N_Shore	− 927	*C3orf80*	0.00278	1.01E−47	+++++ − +++++++++
cg18183624	17	47,076,904	*IGF2BP1*	S_Shore	0	*IGF2BP1*	0.00277	8.36E−80	++++++++++++++
cg04213841	13	49,792,685	*NA*	N_Shore	− 1788	*MLNR*	0.00245	3.60E−43	+++++?+++++++++
cg07738730	17	47,077,165	*IGF2BP1*	S_Shore	0	*IGF2BP1*	0.00217	2.87E−65	++++++++++++ − +
cg09476997	16	2,087,932	*SLC9A3R2*	N_Shore	0	*SLC9A3R2*	0.00208	2.41E−49	++++++++++++++
cg04347477	12	1.25E+ 08	*NCOR2*	Island	833	*NCOR2*	−0.00361	3.38E−32	----------
cg08943494	11	36,422,615	*PRR5L*	OpenSea	69	*PRR5L*	−0.00360	1.95E−24	----------
cg20334115	1	2.26E+ 08	*PYCR2*	N_Shelf	0	*PYCR2*	−0.00350	1.40E−35	----------
cg16725984	16	89,735,184	*C16orf55*	Island	0	*C16orf55*	−0.00325	3.70E−26	----------
cg16103712	8	99,023,869	*MATN2*	OpenSea	7355	*MATN2*	−0.00304	2.70E−129	----------

*Coefficient corresponding to methylation change per additional day of gestational age
**Order of included cohorts in the meta-analysis: MoBa1, MoBa2, MoBa3, EDEN, EXPOSOMICS (Environage+PiccoliPlus+RHEA), CHS, IOWF2, Generation R, Project Viva, CBC (Hispanic), CBC (White), ALSPAC, PREDO, CHAMACOS and INMA."?" Means that CpG was not measured in that cohort

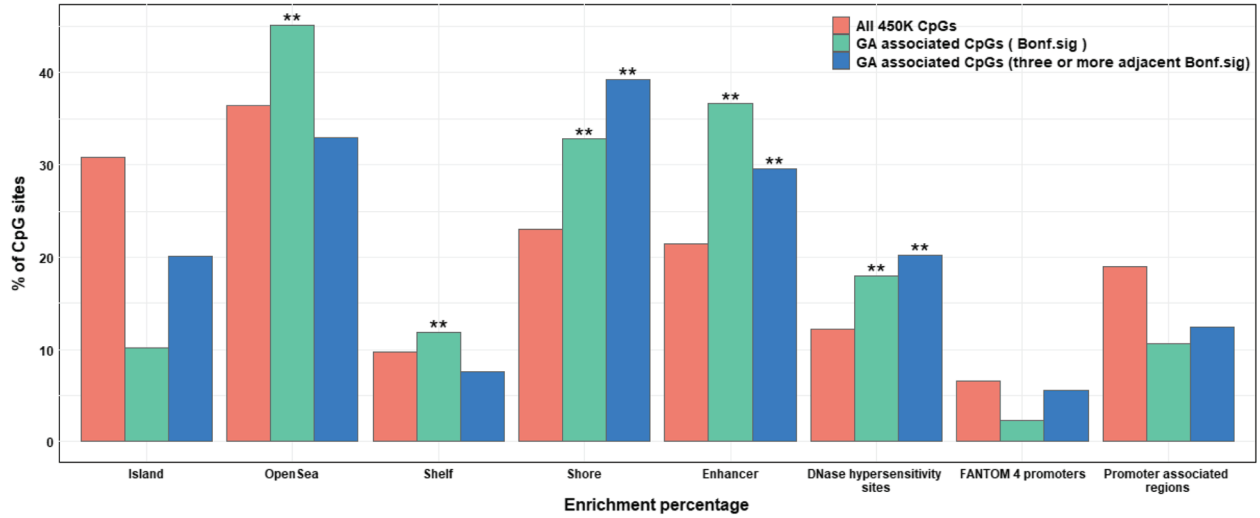


Fig. 3 Position enrichment analyses for CpGs. Salmon: all CpGs in the Illumina450k annotation file, green: CpGs significantly associated with GA after Bonferroni correction ($P < 1.06 \times 10^{-7}$) and blue: three or more adjacent CpGs associated with GA after Bonferroni correction ($P < 1.06 \times 10^{-7}$). "**" represent significant two-sided doubling mid *P* value of the hypergeometric test

included; GEN3G) ($n = 3593$). We identified 5930 sites significantly associated with gestational age at Bonferroni correction ($P < 1.06 \times 10^{-7}$, median difference in mean methylation per additional gestational week = 0.43%, IQR = [0.32%–0.58%]). The vast majority (5399; 91%) of these differentially methylated CpGs overlapped with those found in the main analyses (no complications model) without exclusion of those born preterm (Fig. 4).

Selection of CpGs for downstream analyses

Given the large number of significant associations in our main model (8899 CpGs), we focused subsequent analyses on loci including at least three adjacent CpGs that survived Bonferroni correction [43]. There were 1276 differentially methylated CpGs in 325 unique genes that fulfilled this criterion (Additional file 1: Table S5). As in the overall data, we observed a slight predominance of negative ($n = 702$; 55%) versus positive ($n = 574$; 45%) directions of effect (Fig. 2a). The lowest *P* value, $P = 1.2 \times 10^{-93}$, was observed for cg04276536 (*CCDC102A*, chromosome 16). As for the full EWAS results, the largest negative and positive association effect sizes were observed for cg04347477 (*NCOR2*) and cg13036381 (*LOC401097*), respectively. These 1276 CpGs had the same CpG localization enrichment pattern as the full set of Bonferroni-significant CpGs ($n = 8899$), except that there was a relative depletion in CpG island shelves (7.6% versus 10% overall, $P_{enrichment} = 2.3 \times 10^{-12}$) and open sea (32% versus 37%, $P_{enrichment} = 2.4 \times 10^{-12}$) (Fig. 3).

Differentially methylated region (DMR) analyses

Using two different methods for DMR analysis of gestational age in relation to newborn DNA methylation, we identified 4479 significant (Šidák-corrected $P < 0.01$) DMRs from the comb-p method and 14,671 significant (FDR $P < 0.01$) DMRs from DMRcate, respectively, including 2375 DMRs (representing 11,861 CpGs) that were significant based on both approaches (Additional file 1: Table S6). Out of the 8899 Bonferroni significant single CpGs, 2289 CpGs overlapped with CpGs in identified in the combined DMR analyses (11,861 CpGs). Moreover, from loci included by the three or more adjacent CpG selection ($n = 1276$), 521 CpGs overlapped with those identified in the combined DMR analyses. Of note, out of the 1276 CpGs, 1223 and 1231 CpGs were captured by DMRs identified using the comb-p and DMRcate independent approaches, respectively.

Assessment of CpG methylation in earlier embryonic stages

We examined whether the CpGs detected in cord blood (that originate from embryonic germ layer mesoderm) were differentially methylated in relation to gestational age in other fetal tissues, lung and brain that originate from the two other embryonic germ layers, ectoderm and endoderm, respectively, collected prenatally [47, 48]. To this end, we performed look-up analyses in DNA methylation data for 74 fetal lung samples representing gestational age 59 to 122 days (~ 8 to 17 completed gestational weeks) [47]. Out of the 1276 CpGs, selected based on three or more adjacent CpGs from our no complications model, 1030 CpGs were available in the fetal lung dataset. We observed associations at Bonferroni look-up level correction significance (0.05/1030; $P < 4.85 \times 10^{-5}$) between DNA methylation levels in fetal lung tissue and gestational

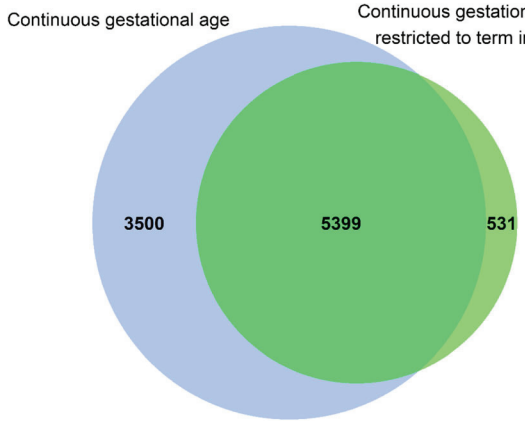


Fig. 4 Overlap between Bonferroni-significant CpG sites from two different analyses after exclusion of maternal and delivery start with induction or caesarean section ("no complication" model). The blue colour represents the continuous gestational age main model, and the green represents the continuous model restricted to term only. Overlap of findings alters the colour

age at tissue collection for 151 (15%) CpGs (Additional file 1: Table S7). Of these 151 (58 negatively and 93 positively associated), 78 showed the same direction of association with gestational age in cord blood and fetal lung tissue. The look-up analyses of fetal brain tissue were undertaken in 179 samples representing 23 to 184 days (~ 3 to 26 completed weeks) [48]. Out of the 1276 CpGs, we found significant associations (using Bonferroni correction $P < 1.06 \times 10^{-7}$ cut-off since only this data was available for analyses; Additional file 1: Table S8) for 268 CpGs (21%) in relation to gestational age at tissue collection. Of these 268 sites, 227 had same direction of effect in the cord blood and fetal brain data. We found enrichment more than expected by chance for our cord blood gestational age associated CpGs ($n = 1276$) in fetal lung ($P = 2.1 \times 10^{-4}$) and brain ($P = 3.9 \times 10^{-57}$) tissue. Thirty CpGs showed significant associations with gestational age in all three tissues (cord blood, fetal lung and fetal brain).

Assessment of CpG methylation in older children

We examined whether the differentially methylated CpGs detected in cord blood samples were associated with gestational age at birth in whole blood from older children. We conducted three separate meta-analyses (no complications model) reflecting different age periods in a total of 2481 children: (i) Early childhood (4–5 years; $n = 453$ from 4 cohorts); (ii) school age (7–9 years; $n = 899$ from 5 cohorts) and (iii) adolescence (16–18 years; $n = 1129$ from 5 cohorts), Additional file 1: Table S1. Of the 1276 three or more adjacent genome-wide significant CpGs from our analyses in cord blood, 1258 CpGs were available for analyses in all older age groups. Out of these CpGs, we observed 40 sites in early childhood, 60 sites in school age, and 60 sites in adolescence to be associated with gestational age at the nominal significance level, $P < 0.05$ with the same direction of effect (Additional file 1: Table S9). However, no CpG survived Bonferroni look-up level correction (0.05/1258; $P < 3.97 \times 10^{-5}$). One CpG (cg26385222 annotated to *TMEM176B*) previously associated with gestational age at birth [27] was nominally significant in all age groups with same direction of effect.

Longitudinal analysis

The results of the longitudinal analyses of blood DNA methylation in the INMA Study ($n = 177$ with paired samples from birth and 4 years) and the ALSPAC Study ($n = 281$ with samples collected at birth, 7 and 17 years) are provided in Additional file 1: Table S10. The vast majority of gestational age associated CpGs ($n = 1054/1276$; 83%) underwent changes in methylation levels with age. Both increasing and decreasing patterns of change during early childhood (4 years) were observed, followed by stabilization during school age (7 years). For example, for cg08943494 in *PRR5L* on chr 11, an initial level of 61.5% and 51.4% in cord blood DNA methylation in INMA and ALSPAC respectively, decreased by 8.2% per year on average during early childhood in INMA and by 3.3% per year on average up to school age in ALSPAC, but then negligible further changes were seen from 7 to 17 years (Fig. 5A). In contrast, increasing levels were seen for cg18183624 (chr 17; *IGF2BP1*), from an initial 48.8% and 38.7% in cord blood DNA methylation in INMA and ALSPAC, respectively, with a 5.1% per year on average between birth to 4 years in INMA and 1.9% per year on average between birth to 7 years, but after that no changes from 7 to 17 years. (Fig. 5B).

Of the 1054 CpGs displaying changes in DNA methylation levels with age, there were 589 CpGs where gestational age was associated with changes in DNA methylation levels (i.e. where an interaction between gestational age and age was found) from birth to 4 years (INMA) and 460 CpGs with changes from birth to 7 years (ALSPAC). However, only 30 of the 1054 CpGs changed significantly in DNA methylation between 7 and 17 years (ALSPAC), suggesting that gestational age-related changes in DNA methylation levels had largely stabilized by age 7.

We identified 222 stable CpGs out of 1276 (17%) that did not change appreciably from birth to adolescence. As an example, the stable DNA methylation at cg27058497 (*RUNX3*, chromosome 1) is shown in Fig. 5C. A much lower proportion of the gestational age associated CpGs were stable from

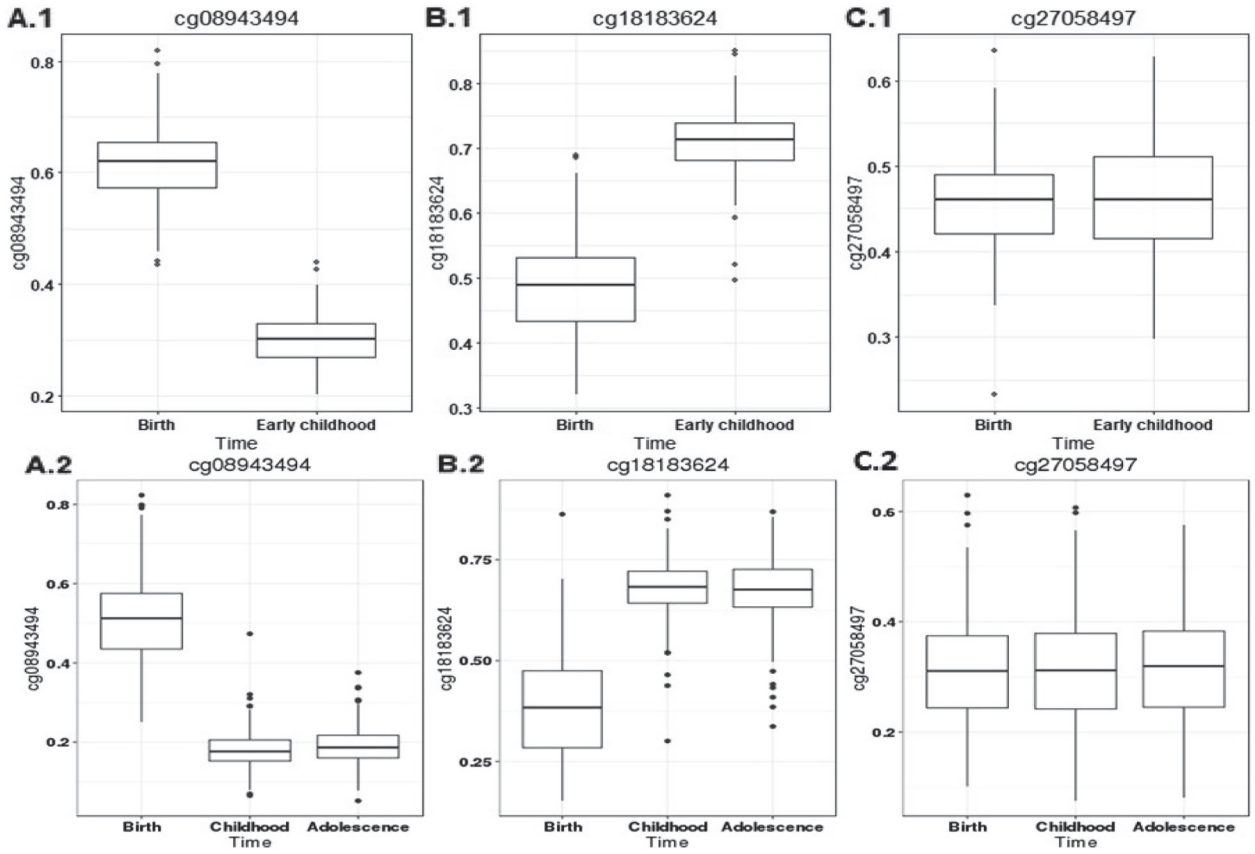


Fig. 5 Change in DNA methylation during childhood and adolescence for selected CpG sites associated with gestational age. **A** Decreasing methylation levels from birth to childhood (**A.1**) and stabilization during adolescence (**A.2**). **B** Increasing methylation levels from birth to childhood and stabilization during adolescence. **C** Stable CpGs that did not change during childhood or adolescence; (1) INMA from birth to early childhood and (2) ALSPAC from birth to adolescence. The figures show representative single CpGs for each category (**A–C**)

birth to adolescence compared to all CpGs on the array (17% versus 71%, $P_{\text{enrichment}} = 2.23\times 10^{-308}$).

Enrichment for biological processes and pathways

Using the complete list of 8899 CpGs annotated to 4966 genes, these were enriched for 1784 GO terms including regulation of cellular and biological processes, system development, different signaling pathways and organ development (Additional file 1: Table S11). Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses revealed 124 significant terms at FDR < 0.05 representing a variety of human diseases, most notably various cancers, viral infections, metabolic processes and immune-related disorders (Additional file 1: Table S12). The 325 genes annotated to the 1276 CpGs, selected by virtue of three or more CpGs being localized to the same gene, were enriched for 198 Gene Ontology (GO) terms very similar to those identified using Bonferroni significant CpGs (Additional file 1: Table S13). When restricting analyses to the 222 longitudinally stable CpGs, corresponding to 139 genes, 13 significant KEGG terms were revealed, primarily representing infection- and immune-related disorders (Additional file 1: Table S14). For 186 genes annotated to the 1054 CpGs changing with postnatal age, only one KEGG terms were identified as statistically significant ($P = 1.2 \times 10^{-3}$ for the term MAPK signaling pathways; Additional file 1: Table S14).

Correlation of DNA methylation and gene expression

For the 1276 CpGs differentially methylated in relation to gestational age with at least 3 adjacent CpGs, we assessed correlations between DNA methylation and gene expression (*cis*-eQTMs). From a publicly available dataset of expression and DNA methylation measured in 38 cord blood samples [51–53], 1174 out of the 1276 CpGs were located within a 500-kb (+/− 250 kb) window of a transcript cluster. Of these 1174, 246 unique CpGs (367 total CpG-transcript associations) correlated significantly with gene expression (Bonferroni $P < 0.05$, Additional file 1: Table S15). Forty-six percent of these DNA methylation-expression correlations were negative, with the lowest $P = 3.55 \times 10^{-6}$ coeff = − 6.03 for cg01332054 and *SEMA7A* expression and the largest negative effect estimate (− 12.69) for cg26179948 and *JAZF1* expression

(Additional file 3: Figure S3 A, B). Fifty-four percent were positive, with the lowest $P = 1.04 \times 10^{-5}$ coeff = 2.88 for cg20139800 and *MOG* expression and the largest positive effect estimate (19.35) for cg03665259 and *CDSN* expression (Additional file 3: Figure S3 C, D).

Discussion

In this large consortium-based meta-analysis, we identified 8899 sites across the genome where gestational age at birth was associated with cord blood DNA methylation. We also identified numerous unique differentially methylated regions (DMRs) associated with gestational age by applying two independent methods. The results were consistent when restricted to births at term, demonstrating that the majority of our results were not driven by preterm births. We confirmed many of the findings from previously published EWAS of gestational age [23, 26, 27, 29, 30, 67] and found a very high correlation between the significant CpG point estimates in previously published datasets compared to our study (e.g. corr = 0.92 between Hannon et al. CpGs and our data; Additional file 1: Table S16), but importantly, we also found 3343 CpGs corresponding to 2577 genes that had not been described previously. There was a general lack of stability of the cord blood findings into childhood and adolescence. However, there was a significant overlap of differentially methylated CpGs in cord blood, fetal brain and lung tissues.

We found that various functional elements were enriched among gestational age-associated CpGs. CpG island shores, enhancers and DNase I hypersensitive sites were particularly susceptible to DNA methylation changes in relation to gestational age, suggesting that these differentially methylated sites are of functional importance [68].

We found clear overlap of differentially methylated CpGs in cord blood, fetal brain and fetal lung tissues in relation to gestational age. Thus, our cord blood findings seem to partly capture the epigenomic plasticity of prenatal development across tissues. The gene with the largest negative magnitude of association with cord blood DNA methylation in relation to gestational age, *NCOR2*, was also differentially methylated in brain and lung fetal tissues. *NCOR2* is involved in vitamin A metabolism and has previously been associated in GWAS with lung function [69]. Vitamin A supplementation is suggested to reduce the risk of bronchopulmonary dysplasia in extremely preterm-born children [70]. Differential methylation of *NCOR2* in neurons associated with ageing has been reported [71]. The gene with the second largest magnitude of negative association with methylation at birth, *PRR5L*, has been linked in GWAS to allergic diseases, found downregulated (expression) in osteoarthritis, and differentially methylated in type II diabetes [72–74]. The gene with the lowest *P* value in our EWAS, *MATN2* plays a critical role in the differentiation and maintenance of skeletal muscles, peripheral nerves, liver and skin during development and regeneration [75] and is suggested as a potential biomarker in the early stage of osteoarthritis [76].

Differentially methylated CpGs associated with gestational age in cord blood were also present in our childhood and adolescence analyses. The only CpG (cg26385222, *TMEM176B*) that was associated with gestational age at all three time points (birth, childhood and adolescence) has been associated with gestational age in cord blood in previous studies [27]. The protein encoded by *TMEM176B* has also been suggested as a potential biomarker for various cancers [77]. The low number of significant associations with gestational age at older ages with no CpG surviving multiple test correction may be partially explained by smaller sample sizes in childhood and adolescence than at birth and by the fact that many later exposures may obscure the association. However, in agreement with the cross-sectional analyses, our longitudinal analyses showed that DNA methylation at gestational age-associated CpGs typically undergoes dynamic changes during early childhood to a much higher degree than overall for CpGs on the 450K array. For the majority of these dynamics CpGs, change was most prominent during the first years of life, with many sites tending stabilize in methylation levels by school age. We also identified a subset of the CpGs differential methylated at birth (17%) which seem stable over time. For these CpGs, the early alteration of methylation levels by length of gestation was found stable postnatally across childhood and into adolescence.

In recent analyses by Xu et al, 14,150 CpGs related to childhood age were identified [78] and we found 280 overlapping with these CpGs among our 1276 CpG list. Moreover, a study by Acevedo et al. showed 794 age-modified CpGs within 3 to 60 months after birth and 57 CpGs were overlapping with our 1276 CpG list [79]. Thus, a proportion of gestational age-related CpGs are also associated with postnatal ageing. But similar to results from Simpkin et al. [80], we observed very little overlap (only 3 CpGs) with the CpGs used to derive epigenetic age by the Hannum and Horvath approach [81, 82] or the epigenetic clock for gestational age at birth (10 CpGs overlapping) [28]. It should be noted that these studies primarily used the Illumina 27K array for analyses, which makes comparison difficult.

In the functional analyses, we observed significant enrichment for several GO terms related to embryonic development, regulation of process and immune system development. The pathway analyses identified a subset of these genes linked to diseases also associated with low gestational age, for example asthma

[83], inflammatory bowel disease [84], type I/II diabetes [85] and cancer (leukaemia) [86]. Importantly, genes annotated to CpGs found stable across childhood also showed enrichment for infection- and immune-related conditions. Whether cord blood DNA methylation at these CpGs affects later disease risk remains to be studied. Interestingly, differentially methylated loci in relation to asthma development have been recently identified in newborns [87]. The stable CpG cg27058497 (*RUNX3*) has been associated with in utero tobacco smoking exposure [88], childhood asthma [89], oesophagus squamous cell carcinoma [90] and chronic fatigue syndrome [91]. Despite adjustment for maternal smoking in our gestational age EWAS model, we observed overlap between all FDR hits from our gestational age EWAS with those FDR hits presented in the maternal smoking related DNA methylation [20] with an overlap of 2302/47,324 CpGs (4.9%, $P_{\text{enrichment}} < 2.2 \times 10^{-308}$). This overlap likely reflects some pregnant women under reporting their smoking behaviour and the fact that smoking-related CpGs capture quantitative smoking history better than self-report [92, 93]. However, we cannot rule out the possibility that some overlapping CpGs could be involved in biologic pathways linking smoking to the well-established consequence of shorter gestational length [94]. Other potential confounders not accounted for in this study such as maternal obesity and alcohol intake may influence offspring DNA methylation although we have found in the PACE consortium that their impact on methylation [95, 96] is very modest compared with maternal smoking in pregnancy which was included in our models.

This paper aimed at identifying CpGs associated with gestational age while adjusting for birth weight. In a recent PACE paper, we found 1071 CpGs at Bonferroni significant levels association with birth weight [97]. Even after adjustment of birth weight in our gestational age EWAS, we observed overlap between the birth weight EWAS and the current gestational age EWAS for 373/1071 CpGs (34.9% $P_{\text{enrichment}} < 2.2 \times 10^{-308}$). These two perinatal factors, birth weight and gestational age, may have a shared impact on DNA methylation in newborns. However, it is difficult to disentangle the effects of these correlated factors.

To further investigate a potential functional impact of our differentially methylated CpGs, we examined correlations with gene expression in cord blood. We found multiple *cis*-eQTMs among the gestational age-related CpGs where methylation was strongly correlated with gene expression in cord blood, implying that the identified CpGs may have a direct functional effect in newborns. *IGF2BP1*, known to be involved in adiposity and cardiometabolic disease risk [98], and to play an essential role in embryogenesis and carcinogenesis [99, 100], was the most significant positively differentially methylated CpG in cord blood. Low gestational age is a well-established risk factor for later cardiometabolic disease [101]. Our expression findings likely reflect relevant for health outcomes associated with low gestational age.

There are potential study limitations in our study including heterogeneity in normalization and quality control (QC) protocols since individual cohorts performed their own QC and normalization. However, one of our previous EWAS meta-analysis reported robust results comparing the non-normalized methylation and different data processing methods used across the cohorts for normalization [20]. Furthermore, between-study heterogeneity at our pre-specified threshold was observed for only a minority of differentially methylated CpGs. Cohorts collected gestational age data from medical records, birth certificates or questionnaires in two ways, either ultrasound estimates and/or according to last menstrual period (or combined estimates), which may introduce bias. However, gestational age determined by ultrasound correlates well with last menstrual period data [102]. Despite a large sample size, we had few extreme premature births included in our dataset. Interpretation of effects of DNA methylation on gene expression was done for *cis*-effects only, not *trans*-effects. Since our analyses were primarily cross-sectional, we cannot infer the temporality in the associations and we cannot assume associations are causal [103]. We recognize the possibility that the observed methylation patterns represent fetal maturity, accompanying a "normal" developmental process or determining time in utero; it was however not possible to include foetuses who did not survive pregnancy most of whom will have been delivered very early. The majority of study participants were of European ancestry, and very few cohorts were Hispanic. We were unable to explore ethnic differences in detail since that would require large sample sizes for each ethnic group. However, when analyses were restricted to European-ancestry cohorts, the results were essentially identical with correlation coefficient 0.97 (Additional file 3: Figure S4) to those with all cohorts included. Finally, we acknowledge a potential limitation by applying a filter (regions with at least three or more adjacent CpGs with a Bonferroni-corrected P value < 0.05) in order to capture a set of genes robustly affected by gestational age, which may have led to potentially important single CpGs not being included in the functional analyses. In addition, genes with few CpGs represented on the 450K array are likely underrepresented in the downstream analyses. The strengths of our study are large sample size, the comprehensive analyses using robust statistical methods, as well as the

availability of samples at multiple ages and our ability to compare our findings with those in fetal tissue datasets. To account for potential cell type effects, we adjusted our models for estimated cell counts using cord blood and adult whole blood references [35, 36]. However, we acknowledge the limitations of available blood cell type reference data sets and recognize that some of the signals we identified as effects of gestational age might reflect differences in cell type composition that we did not completely control. Larger panels that better capture cell type composition across the range of gestational age would be a useful advance. Although we present data on all available participants in our all births model, we based our study conclusions on the main no complication model results, after excluding samples related to delivery induced by medical interventions (induction and/or caesarean section) and maternal complications.

Conclusions

We show that DNA methylation at numerous CpG sites and DMRs across the genome is associated with gestational age at birth. Our results provide a comprehensive catalogue of differential methylation in relation to this important factor, which may serve as utility to the growing community of researchers studying the developmental origins of adult disease. Identified CpGs were linked to multiple functional pathways related to human diseases and enriched for several categories of biological processes critical to fetal development. As such, many sites might capture epigenomic plasticity of fetal development across tissues. We also found that blood DNA methylation levels in identified CpGs change over time for a majority of CpGs and that levels stabilize after school age. Taken together, our findings provide new insight into epigenetics related to preterm birth and gestational age.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13073-020-0716-9.

Additional file 1: Table S1. Cohort-specific results from epigenome-wide association analyses of gestational age. **Table S2.** Normalization technique and phenotype definitions used by each cohort. **Table S3.** Bonferroni-significant CpGs from the meta-analysis on the association between continuous gestational age (no complications model) and offspring DNA methylation at birth adjusted for estimated cell counts. **Table S4.** Bonferroni-significant CpGs from the meta-analysis on the association between continuous gestational age (all births model) and offspring DNA methylation at birth adjusted for estimated cell counts. **Table S5.** Gene regions that had at least three consecutive Bonferroni significant CpG sites from the continuous gestational age analyses (no complications model). **Table S6.** DMRs ($n = 2375$) for gestational age in relation to newborn methylation (no complication model) identified by using both comb-p ($P < 0.01$) and DMRcate (FDR < 0.01) methods. **Table S7.** DNA methylation analyses in fetal lung tissue using the no complication gestational age three or more consecutive CpG list. **Table S8.** DNA methylation analyses in fetal brain tissue using the no complication gestational age three or more consecutive CpG list. **Table S9.** Methylation look-up analyses in older children using the no complication gestational age three or more consecutive CpG list. **Table S10.** Longitudinal analysis of methylation levels in the INMA and ALSPAC studies using the no complication gestational age three or more consecutive CpG list. **Table S11.** Gene Ontology (GO) term enrichment analyses for bonferroni-significant CpGs from the meta-analysis (no complications model). **Table S12.** KEGG pathway analyses for bonferroni-significant CpGs from the meta-analysis (no complications model). **Table S13.** Gene Ontology (GO) term enrichment analyses for three or more CpGs being localized to the same gene. **Table S14.** KEGG pathway analyses for stable and dynamic CpGs. **Table S15.** Correlation between methylation and gene expression levels in cord blood (cis-effects). **Table S16.** The replication of bonferroni-significant CpGs from the meta-analysis (no complications model) in previous publication.

Additional file 2. Supplementary information.

Additional file 3: Figure S1. Forest plot for the top 10 Bonferroni-significant CpGs from the meta-analysis on the association between continuous GA and offspring DNA methylation at birth adjusted for estimated cell proportions. **Figure S2.** Sensitivity analysis: Correlation of the point estimates for the no complications model main association of DNA methylation with gestational age (y-axis representing 3648 participants from 17 cohorts) with point estimates for a meta-analysis after excluding three cohorts (MoBa1, MoBa2 and ALSPAC) that were included in a previous publication1,2 (x-axis representing 2190 participants from 14 cohorts). **Figure S3.** Correlations between methylation and gene expression levels for selected four pairs. First, we created residuals for mRNA expression and residuals for DNA methylation and used linear regression models to evaluate correlations between expression residuals and methylation residuals. These residual models were adjusted for covariates, estimated white blood cell proportions, and technical variation. **Figure S4.** Sensitivity analysis: Correlation of the point estimates for the no complications model main association of DNA methylation with gestational age (y-axis representing 3648 participants from 17 cohorts) with point estimates for a meta-analysis after excluding Non-European three cohorts (CBC, CHS and CHAMACOS) (x-axis representing 3290 participants from 14 cohorts).

Acknowledgements

For all studies, detailed information can be found in Additional file 2: Supplementary information.

Funding

This study was specifically funded by a grant from the European Research Council (TRIBAL, grant agreement 757919). For all studies, detailed information can be found in Additional file 2: Supplementary information. Open access funding provided by Uppsala University.

Availability of data and materials

Genome-wide DNA methylation meta-analysis summary statistics corresponding to the main analysis presented in this manuscript are available at figshare (https://doi.org/10.6084/m9.figshare.11688762.v1) [44]. Individual cohort level data may be available by application to the relevant institutions after obtaining required approvals. All datasets used are previously published as described in Felix et al. [31]. Additional details and references to the study cohorts are available in Additional file 2.

Authors' contributions

EM and SJL conceived and designed the study with input from the project group (SKM, GHK, JF, M-FH, AG, NH, MW, OS, PB, JK, SER, C-JX, AC, OG, CAM, CS, AK and LKK). GCS (ALSPAC and GOYA), SKM (BAMSE, EDEN and PIAMA), RR (CBC), OS (CHAMACOS), LG (CHS), PJ (EXPOSOMICS: Environage, Piccoli-Plus and RHEA), LKK (GECKO), CA (Gen3G), FOV (Generation R), LAS (INMA), FIR (IOW F1), HZ (IOW F2), SER (MoBa1 and MoBa2), AN (MoBa3), MW (NFBC86), DC (PREDO), AC (Project Viva) and PEM (Raine) conducted the cohort-specific analyses. Longitudinal analyses were performed by SKM (INMA, with support from MB) and GSC (ALSPAC). ATK performed analyses on fetal lung data sets. SKM meta-analyses all results with AN as shadow analyst. SKM performed expression and DNA methylation follow-up analyses and

bioinformatics analysis. SKM, EM and SJL wrote the first draft of the manuscript. All authors (SKM, AN, GCS, LKK, ATK, RR, LG, IAM, PJ, MP, MK, CA, FOV, NK, LAS, FIR, HZ, SS, DC, SLR-S, PEM, DAL, GP, CVB, KH, NB, LG, TSN, EC, PP, LD, EAN, MB, SLE, WK, SZ, CMP, ZH, M-RJ, JL, AAB, DA, PK, CLR, AB, BE, MHS, PV, HS, LB, VWJ, TIAS, MV, SHA, JWH, SEH, PM, TD, EBB, DLD, JMV, JN, KGT, IK, JLW, BH, JS, WN, MCM-K, KR, EO, R-CH, STW, JMA, JB, AK, CS, CA, AC, OG, C-JX, SER, JK, PB, OS, MW,NH, AG, M-FH, JFF, GHK, SJL, EM) read and critically revised subsequent drafts, and approved the final version. Correspondence and material requests should be addressed to EM (erik.melen@ki.se).

Ethics approval and consent to participate

All cohorts acquired ethics approval and informed consent from participants prior to data collection through local ethics committees; detailed information for each cohort can be found in Additional file 2: Supplementary information. Our research conformed to the principles of the Helsinki Declaration.

Consent for publication

Not applicable.

Competing interests

DA Lawlor declares grants from Medtronic Ltd. and Roche Diagnostics and EBB; A Ghantous is identified as personnel of the IARC, the author alone is responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the IARC. The remaining authors declare that they have no competing interests.

Author details

[1]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [2]Department of Clinical Sciences and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [3]Epigenetics Group, International Agency for Research on Cancer, Lyon, France. [4]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [5]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK. [6]Division of Human Nutrition and Health, Wageningen University & Research, Wageningen, the Netherlands. [7]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands. [8]Computational Health Informatics Program, Boston Children's Hospital and Harvard Medical School, Boston, MA, USA. [9]Computational Biology And Informatics, University of California, San Francisco, San Francisco, CA, USA. [10]HDF Comprehensive Cancer Center, University of California, San Francisco, San Francisco, CA, USA. [11]Department of Preventive Medicine, University of Southern California, Los Angeles, USA. [12]Sorbonne Université and INSERM, Epidemiology of Allergic and Respiratory Diseases Department (EPAR), Pierre Louis Institute of Epidemiology and Public Health (IPLESP UMRS 1136), Saint-Antoine Medical School, Paris, France. [13]NIHR-Health Protection Research Unit, Respiratory Infections and Immunity, Imperial College London, London, UK. [14]Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, UK. [15]Centre for Environmental Sciences, Hasselt University, Hasselt, Belgium. [16]ISGlobal, Barcelona Institute for Global Health, Barcelona, Spain. [17]Universitat Pompeu Fabra (UPF), Barcelona, Spain. [18]CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [19]IMIM (Hospital del Mar Medical Research Institute), Barcelona, Spain. [20]Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke (CHUS), Sherbrooke, QC, Canada. [21]The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands. [22]Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands. [23]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Lebanon, USA. [24]School of Water, Energy and Environment, Cranfield University, Cranfield, Bedfordshire MK43 0AL, UK. [25]Division of Epidemiology, Biostatistics, and Environmental Health, School of Public Health, University of Memphis, Memphis, USA. [26]Center for Life Course Health Research, Faculty of Medicine, University of Oulu, Oulu, Finland. [27]Biocenter Oulu, University of Oulu, Oulu, Finland. [28]Department of Genomic of Complex diseases, School of Public Health, Imperial College London, London, UK. [29]Department of Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany. [30]Division of Chronic Disease Research Across the Lifecourse (CoRAL), Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA. [31]School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University, Bentley, Australia. [32]Curtin/UWA Centre for Genetic Origins of Health and Disease, School of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Western Australia, Perth, Australia. [33]Bristol NIHR Biomedical Research Centre, Bristol, UK. [34]Centre for Occupational and Environmental Medicine, Stockholm, Stockholm Region, Sweden. [35]Children's Environmental Health Laboratory, University of California, Berkeley, Berkeley, CA, USA. [36]Division of Neonatology and Pediatrics, Ospedale Versilia, Viareggio, AUSL Toscana Nord Ovest, Pisa, Italy. [37]Department of Public Health & Primary Care, Leuven University, Leuven, Belgium. [38]Department of Medicine, Université de Sherbrooke, Sherbrooke, Canada. [39]Research Unit for Gynaecology and Obstetrics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark. [40]College of Veterinary Medicine, Michigan State University, East Lansing, MI, USA. [41]Department of Health and Human Services, National Institute of Environmental Health Sciences, National Institutes of Health, RTP, Durham, NC, USA. [42]Norwegian Institute of Public Health, Oslo, Norway. [43]Department of Epidemiology and Biostatistics, MRC–PHE Centre for Environment & Health, School of Public Health, Imperial College London, London, UK. [44]Unit of Primary Care, Oulu University Hospital, Oulu, Finland. [45]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland. [46]Turku Institute for Advanced Studies, University of Turku, Turku, Finland. [47]Department of Environmental Health Sciences, Mailman School of Public Health, Columbia University Medical Center, New York, NY, USA. [48]Telethon Kids Institute, University of Western Australia, Perth, Australia. [49]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [50]Center for Environmental Research and Children's Health (CERCH), University of California, Berkeley, Berkeley, CA, USA. [51]MRC-PHE Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [52]Department of Biochemistry, Université de Sherbrooke, Sherbrooke, QC, Canada. [53]Department of medical biology, CIUSSS-SLSJ, Saguenay, QC, Canada. [54]Novo Nordisk Foundation Center for Basic Metabolic Research, Section on Metabolic Genetics, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [55]Department of Public Health, Section of Epidemiology, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [56]Clinical & Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. [57]The David Hide Asthma and Allergy Research Centre, Newport, Isle of Wight, UK. [58]Human Development & Health, Faculty of Medicine, University of Southampton, Southampton, UK. [59]Nuffield Department of Women's and Reproductive Health, University of Oxford, Oxford, UK. [60]Murdoch Children's Research Institute, Australia Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Melbourne, Australia. [61]Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, USA. [62]University of Groningen, University Medical Center Groningen, Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, The Netherlands. [63]Faculty of Health and Medical Sciences, UWA Medical School, University of Western Australia, Perth, Australia. [64]Sachs' Children's Hospital, Södersjukhuset, 118 83 Stockholm, Sweden. [65]Center for Genetic Epidemiology, University of Southern California, Los Angeles, USA. [66]INSERM, UMR1153 Epidemiology and Biostatistics Sorbonne Paris Cité Center (CRESS), Research Team on Early life Origins of Health (EarOH), Paris Descartes University, Paris, France. [67]Department of Pediatric Oncology and Hematology, Oslo University Hospital, Oslo, Norway. [68]University Hospital, Montpellier, France. [69]Department of Dermatology, Charité, Berlin, Germany. [70]University of Basel, Basel, Switzerland. [71]Swiss Tropical and Public Health Institute, Basel, Switzerland. [72]Department of Women's and Children's Health, Karolinska Institutet, Stockholm, Sweden. [73]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [74]Pediatric Allergy and Pulmonology Unit at Astrid Lindgren Children's Hospital, Karolinska University Hospital, Stockholm, Sweden. [75]Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [76]University of Groningen, University Medical Center Groningen, Department of Pediatric Pulmonology and Pediatric Allergology, Beatrix Children's Hospital, GRIAC Research Institute Groningen, Groningen, The Netherlands. [77]Department of Biosciences and Nutrition, Karolinska Institutet, Huddinge, Sweden. [78]Folkhälsa Research Institute, Helsinki, and Stem Cells and Metabolism Research Program, University of Helsinki Finland, Helsinki, Finland. [79]Department of Newborn Medicine, Karolinska University Hospital,

Stockholm, Sweden. [80]Science for Life Laboratory, Stockholm, Sweden. [81]Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA. [82]Sachs' Children's Hospital, South General Hospital, Stockholm, Sweden.

Received: 27 June 2019 Accepted: 30 January 2020

References

1. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet (London, England). 2008;371:75–84.
2. Engle WA. Morbidity and mortality in late preterm and early term newborns: a continuum. Clin Perinatol. 2011;38:493–516.
3. Leung JY, Lam HS, Leung GM, Schooling CM. Gestational age, birthweight for gestational age, and childhood hospitalisations for asthma and other wheezing disorders. Paediatr Perinat Epidemiol. 2016;30:149–59.
4. Raby BA, et al. Low-normal gestational age as a predictor of asthma at 6 years of age. Pediatrics. 2004;114:e327–32.
5. Been JV, et al. Preterm birth and childhood wheezing disorders: a systematic review and meta-analysis. PLoS Med. 2014;11:e1001596.
6. den Dekker HT, et al. Early growth characteristics and the risk of reduced lung function and asthma: a meta-analysis of 25,000 children. J Allergy Clin Immunol. 2016;137:1026–35.
7. Parets SE, Bedient CE, Menon R, Smith AK. Preterm birth and its long-term effects: methylation to mechanisms. Biology. 2014;3:498–513.
8. Kwinta P, Pietrzyk JJ. Preterm birth and respiratory disease in later life. Expert Rev Respir Med. 2010;4:593–604.
9. Hille ET, et al. Functional outcomes and participation in young adulthood for very preterm and very low birth weight infants: the Dutch project on preterm and small for gestational age infants at 19 years of age. Pediatrics. 2007;120:e587–95.
10. Geldof CJ, van Wassenaer AG, de Kieviet JF, Kok JH, Oosterlaan J. Visual perception and visual-motor integration in very preterm and/or very low birth weight children: a meta-analysis. Res Dev Disabil. 2012;33:726–36.
11. Kerkhof GF, Breukhoven PE, Leunissen RW, Willemsen RH, Hokken-Koelega AC. Does preterm birth influence cardiovascular risk in early adulthood? J Pediatr. 2012;161:390–6.e391.
12. Aarnoudse-Moens CS, Weisglas-Kuperus N, van Goudoever JB, Oosterlaan J. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. Pediatrics. 2009;124:717–28.
13. Thunqvist P, et al. Lung function at 8 and 16 years after moderate-to-late preterm birth: a prospective cohort study. Pediatrics. 2016;137(4).
14. Ghartey K, et al. Neonatal respiratory morbidity in the early term delivery. Am J Obstet Gynecol. 2012;207:292.e291–294.
15. Noble KG, Fifer WP, Rauh VA, Nomura Y, Andrews HF. Academic achievement varies with gestational age among children born at term. Pediatrics. 2012;130:e257–64.
16. Talge NM, Allswede DM, Holzman C. Gestational age at term, delivery circumstance, and their association with childhood attention deficit hyperactivity disorder symptoms. Paediatr Perinat Epidemiol. 2016;30:171–80.
17. Yang S, Bergvall N, Cnattingius S, Kramer MS. Gestational age differences in health and development among young Swedish men born at term. Int J Epidemiol. 2010;39:1240–9.
18. Gruzieva O, et al. Epigenome-wide meta-analysis of methylation in children related to prenatal NO2 air pollution exposure. Environ Health Perspect. 2017;125:104–10.
19. Joubert BR, et al. Maternal plasma folate impacts differential DNA methylation in an epigenome-wide meta-analysis of newborns. Nat Commun. 2016;7:10577.
20. Joubert BR, et al. DNA methylation in newborns and maternal smoking in pregnancy: genome-wide consortium meta-analysis. Am J Hum Genet. 2016;98:680–96.
21. Gruzieva O, et al. Prenatal particulate air pollution and DNA methylation in newborns: an epigenome-wide meta-analysis. Environ Health Perspect. 2019;127:57012.
22. Pan WH, et al. Exposure to the gut microbiota drives distinct methylome and transcriptome changes in intestinal epithelial cells during postnatal development. Genome Med. 2018;10:27.
23. Cruickshank MN, et al. Analysis of epigenetic changes in survivors of preterm birth reveals the effect of gestational age and evidence for a long term legacy. Genome Med. 2013;5:96.
24. Cutfield WS, Hofman PL, Mitchell M, Morison IM. Could epigenetics play a role in the developmental origins of health and disease? Pediatr Res. 2007;61:68r–75r.
25. Lee H, et al. DNA methylation shows genome-wide association of NFIX, RAPGEF2 and MSRB3 with gestational age at birth. Int J Epidemiol. 2012;41:188–99.
26. Schroeder JW, et al. Neonatal DNA methylation patterns associate with gestational age. Epigenetics. 2011;6:1498–504.
27. Parets SE, et al. Fetal DNA methylation associates with early spontaneous preterm birth and gestational age. PLoS One. 2013;8:e67489.
28. Knight AK, et al. An epigenetic clock for gestational age at birth based on blood methylation data. Genome Biol. 2016;17:206.
29. Simpkin AJ, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. Hum Mol Genet. 2015;24:3752–63.
30. Bohlin J, et al. Prediction of gestational age based on genome-wide differentially methylated regions. Genome Biol. 2016;17:207.
31. Felix JF, et al. Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. Int J Epidemiol. 2018;47:22–23u.
32. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
33. Hoaglin DC, Iglewicz B, Tukey JW. Performance of some resistant rules for outlier labeling. J Am Stat Assoc. 1986;81:991–9.
34. Venables WR, Ripley BD. Modern Applied Statistics with S. New York: Springer-Verlag; 2002.
35. Bakulski KM, et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. Epigenetics. 2016;11:354–62.
36. Reinius LE, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS One. 2012;7:e41361.
37. Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics (Oxford, England). 2014;30:1363–9.
38. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics (Oxford, England). 2010;26:2190–1.
39. Rice K, Higgins JP, Lumley T. A re-evaluation of fixed effect(s) meta-analysis. J R Statist Soc A. 2018;181:205–27.
40. R Core Team. R Foundation for Statistical Computing; Vienna: R: A language and environment for statistical computing; 2013. http://www.R-project.org/.
41. Han B, Eskin E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. Am J Hum Genet. 2011;88:586–98.
42. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21:1539–58.
43. Hannula-Jouppi K, et al. Differentially methylated regions in maternal and paternal uniparental disomy for chromosome 7. Epigenetics. 2014;9:351–65.
44. Merid SK et al. Summary statistics Data sets. figshare. 2020. https://doi.org/10.6084/m9.figshare.11688762.v1.
45. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinformatics (Oxford, England). 2012;28:2986–8.
46. Peters TJ, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin. 2015;8:6.
47. Chhabra D, et al. Fetal lung and placental methylation is associated with in utero nicotine exposure. Epigenetics. 2014;9:1473–84.
48. Spiers H, et al. Methylomic trajectories across human fetal brain development. Genome Res. 2015;25:338–52.
49. Kamburov A, Wierling C, Lehrach H, Herwig R. ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Res. 2009;37:D623–8.
50. Kamburov A, et al. ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acids Res. 2011;39:D712–7.
51. Rojas D, et al. Prenatal arsenic exposure and the epigenome: identifying sites of 5-methylcytosine alterations that predict functional changes in gene expression in newborn cord blood and subsequent birth outcomes. Toxicol Sci. 2015;143:97–106.
52. Rager JE, et al. Prenatal arsenic exposure and the epigenome: altered microRNAs associated with innate and adaptive immune signaling in newborn cord blood. Environ Mol Mutagen. 2014;55:196–208.
53. Barrett T, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41:D991–5.

54. Ma X, et al. Ethnic difference in daycare attendance, early infections, and risk of childhood acute lymphoblastic leukemia. Cancer Epidemiol Biomarkers Prev. 2005;14:1928–34.
55. McConnell R, et al. Traffic, susceptibility, and childhood asthma. Environ Health Perspect. 2006;114:766–72.
56. Eskenazi B, et al. CHAMACOS, a longitudinal birth cohort study: lessons from the fields. J Childrens Health. 2003;1:3–27.
57. Heude B, et al. Cohort profile: the EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. Int J Epidemiol. 2016;45:353–63.
58. Vineis P, et al. The exposome in practice: design of the EXPOsOMICS project. Int J Hyg Environ Health. 2017;220:142–51.
59. Kruithof CJ, et al. The generation R study: biobank update 2015. Eur J Epidemiol. 2014;29:911–27.
60. Guxens M, et al. Cohort profile: the INMA--INfancia y Medio Ambiente--(environment and childhood) project. Int J Epidemiol. 2012;41:930–40.
61. Everson TM, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive random Forest feature selection. Genome Med. 2015;7:89.
62. Girchenko P, et al. Cohort profile: prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. Int J Epidemiol. 2017;46:1380–1381g.
63. Oken E, et al. Cohort profile: project viva. Int J Epidemiol. 2015;44:37–48.
64. Xu CJ, et al. DNA methylation in childhood asthma: an epigenome-wide meta-analysis. Lancet Respir Med. 2018;6:379–88.
65. Jarvelin MR, Hartikainen-Sorri AL, Rantakallio P. Labour induction policy in hospitals of different levels of specialisation. Br J Obstet Gynaecol. 1993;100:310–5.
66. Straker L, et al. Cohort Profile: The Western Australian Pregnancy Cohort (Raine) Study-Generation 2. Int J Epidemiol. 2017;46:1384–1385j.
67. Hannon E, et al. Variable DNA methylation in neonates mediates the association between prenatal smoking and birth weight. Philos Trans Ro Soc Lond. 2019;374:20180120.
68. Ziller MJ, et al. Charting a dynamic DNA methylation landscape of the human genome. Nature. 2013;500:477–81.
69. Minelli C, et al. Association of Forced Vital Capacity with the developmental gene NCOR2. PLoS One. 2016;11:e0147388.
70. Garg BD, Bansal A, Kabra NS. Role of vitamin A supplementation in prevention of bronchopulmonary dysplasia in extremely low birth weight neonates: a systematic review of randomized trials. J Matern Fetal Neonatal Med. 2019;32:2608-15.
71. Gasparoni G, et al. DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. Epigenetics Chromatin. 2018;11:41.
72. Ferreira MAR, et al. Eleven loci with new reproducible genetic associations with allergic disease risk. J Allergy Clin Immunol. 2019;143:691-9.
73. Wang X, Ning Y, Guo X. Integrative meta-analysis of differentially expressed genes in osteoarthritis using microarray technology. Mol Med Rep. 2015;12:3439–45.
74. Al Muftah WA, et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. Clin Epigenetics. 2016;8:13.
75. Korpos E, Deak F, Kiss I. Matrilin-2, an extracellular adaptor protein, is needed for the regeneration of muscle, nerve and other tissues. Neural Regen Res. 2015;10:866–9.
76. Zhang S, et al. Matrilin-2 is a widely distributed extracellular matrix protein and a potential biomarker in the early stage of osteoarthritis in articular cartilage. Biomed Res Int. 2014;2014:986127.
77. Cuajungco MP, et al. Abnormal accumulation of human transmembrane (TMEM)-176A and 176B proteins is associated with cancer pathology. Acta Histochem. 2012;114:705–12.
78. Xu CJ, et al. The emerging landscape of dynamic DNA methylation in early childhood. BMC Genomics. 2017;18:25.
79. Acevedo N, et al. Age-associated DNA methylation changes in immune genes, histone modifiers and chromatin remodeling factors within 5 years after birth in human blood leukocytes. Clin Epigenetics. 2015;7:34.
80. Simpkin AJ, et al. Prenatal and early life influences on epigenetic age in children: a study of mother-offspring pairs from two cohort studies. Hum Mol Genet. 2016;25:191–201.
81. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14:R115.
82. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49:359–67.
83. Goyal NK, Fiks AG, Lorch SA. Association of late-preterm birth with asthma in young children: practice-based study. Pediatrics. 2011;128:e830–8.
84. Sonntag B, et al. Preterm birth but not mode of delivery is associated with an increased risk of developing inflammatory bowel disease later in life. Inflamm Bowel Dis. 2007;13:1385–90.
85. Li S, et al. Preterm birth and risk of type 1 and type 2 diabetes: systematic review and meta-analysis. Obes Rev. 2014;15:804–11.
86. Wang YF, Wu LQ, Liu YN, Bi YY, Wang H. Gestational age and childhood leukemia: A meta-analysis of epidemiologic studies. Hematology (Amsterdam, Netherlands). 2018;23:253–62.
87. Reese SE, et al. Epigenome-wide meta-analysis of DNA methylation and childhood asthma. J Allergy Clin Immunol. 2019;143:2062-74.
88. Maccani JZ, Koestler DC, Houseman EA, Marsit CJ, Kelsey KT. Placental DNA methylation alterations associated with maternal tobacco smoking at the RUNX3 gene are also associated with gestational age. Epigenomics. 2013;5:619–30.
89. Yang IV, et al. DNA methylation and childhood asthma in the inner city. J Allergy Clin Immunol. 2015;136:69–80.
90. Zheng Y, Zhang Y, Huang X, Chen L. Analysis of the RUNX3 gene methylation in serum DNA from esophagus squamous cell carcinoma, gastric and colorectal adenocarcinoma patients. Hepato-gastroenterology. 2011;58:2007–11.
91. de Vega WC, Herrera S, Vernon SD, McGowan PO. Epigenetic modifications and glucocorticoid sensitivity in myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS). BMC Med Genet. 2017;10:11.
92. Reese SE, et al. DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. Environ Health Perspect. 2017;125:760–6.
93. Valeri L, et al. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? Epigenomics. 2017;9:253–65.
94. Warren GW, Alberg AJ, Kraft AS, Cummings KM. The 2014 surgeon General's report: "the health consequences of smoking--50 years of progress": a paradigm shift in cancer care. Cancer. 2014;120:1914–6.
95. Sharp GC, et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: findings from the pregnancy and childhood epigenetics (PACE) consortium. Hum Mol Genet. 2017;26:4067–85.
96. Sharp GC, et al. Maternal alcohol consumption and offspring DNA methylation: findings from six general population-based birth cohorts. Epigenomics. 2018;10:27–42.
97. Kupers LK, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 2019;10:1893.
98. Lu Y, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. Nat Commun. 2016;7:10495.
99. Mahaira LG, et al. IGF2BP1 expression in human mesenchymal stem cells significantly affects their proliferation and is under the epigenetic control of TET1/2 demethylases. Stem Cells Dev. 2014;23:2501–12.
100. Huang X, et al. Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. J Hematol Oncol. 2018;11:88.
101. Cooper R, Atherton K, Power C. Gestational age and risk factors for cardiovascular disease: evidence from the 1958 British birth cohort followed to mid-life. Int J Epidemiol. 2009;38:235–44.
102. Hoffman CS, et al. Comparison of gestational age at birth based on last menstrual period and ultrasound during the first trimester. Paediatr Perinat Epidemiol. 2008;22:587–96.
103. Dyke SOM, et al. Points-to-consider on the return of results in epigenetic research. Genome Med. 2019;11:31.

Publisher's Note

PACE_sex_draft_08132019OS

Meta-analysis of epigenome-wide association studies in newborns and children show widespread sex differences in blood DNA methylation

Olivia Solomon*[1], Karen Huen[1]*, Leanne K. Kupers[2], Matthew Suderman[3], Olena Gruzieva[4], Lu Gao[5], Kelly M. Bakulski[6], Alexei Novoloaca[7], Catherine Allard[8], Irene Pappa[9,10], Paul Yousefi[3,11], Maria Llambrich[12-14], Marta Vives[12,13,15], Dereje D. Jima[16,17], Tuomas Kvist[18], Andrea Baccarelli[19], Gemma C Sharp[3], Brenda Eskenazi[20], Anna Bergström[4], John F. Dou[5], Elena Isaevska[21], Eva Corpeleijn[2], Patrice Perron[8,22], Vincent W.V. Jaddoe[10,23], Ellen Nøhr[24,25], Lea Maitre[12-14], Maria Foraster[12-14], Cathrine Hoyo[26,27], Jari Lahti[18], Dawn L. DeMeo[28],Inger Kull[29,30], Jason I. Feinberg[31], Luigi Gagliardi[32], Luigi Bouchard[33,34], Henning Tiemeier[9,35], Gillian Santorelli[36], Rachel L. Maguire[26,37], Darina Czamara[38], Augusto Litonjua[28], Michelle Plusquin[39], Johanna Lepeule[40], Elisabeth Binder[38,41], Terence Dwyer[42], Ángel Carracedo[43,44], Katri Raikkönen[18], Manolis Kogevinas[12], Tim S. Nawrot[39,45], Monica C. Munthe-Kaas[46,47], Zdenko Herceg[7], Caroline Relton[3], Erik Melén[29,48], Carrie Breton[5], M. Daniele Fallin[31], Akram Ghantous[7], Harold Snieder[2], Marie-France Hivert[22,49,50], Janine F. Felix[10,23], Thorkild I.A. Sørensen[51,52], Juan R González[12-14], Mariona Bustamante[12-14], Susan K. Murphy[37], Emily Oken[49], Stephanie J. London[53], Nina Holland[1]

Affiliations:

[1] Children's Environmental Health Laboratory, Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, USA.
[2] University of Groningen, University Medical Center Groningen, Department of Epidemiology, Groningen, The Netherlands
[3] MRC Integrative Epidemiology Unit, Population Health Sciences, Bristol Medical School, University of Bristol, BS8 2BN, UK
[4] Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
[5] Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90089, USA
[6] School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA
[7] Epigenetics Group, International Agency for Research on Cancer, Lyon, France
[8] Centre de recherche du Centre hospitalier universitaire de Sherbrooke, QC, Canada
[9] Department of Child and Adolescent Psychiatry/ Psychology, Erasmus Medical Center, Sophia Children's Hospital, P.O. Box 2060, 3000 CB, Rotterdam, The Netherlands
[10] The Generation R Study Group, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2060, 3000 CB, Rotterdam, The Netherlands
[11] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[12] ISGlobal, Barcelona Institute for Global Health, Dr Aiguader 88, 08003, Barcelona, Spain
[13] Universitat Pompeu Fabra (UPF), Barcelona, Spain
[14] CIBER Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[15] Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr. Aiguader 88, Barcelona 08003, Spain
[16] Center for Human Health and the Environment, North Carolina State University, Raleigh, NC 27606
[17] Bioinformatics Research Center, North Carolina State University, Raleigh, NC 27606

[18] Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Finland
[19] Laboratory of Precision Environmental Biosciences, Columbia University Mailman School of Public Health, New York, NY, USA
[20] Center for Environmental Research and Children's Health, School of Public Health, University of California, Berkeley, CA, USA.
[21] Cancer Epidemiology Unit, Department of Medical Sciences, University of Turin, Turin, Italy
[22] Department of Medicine, Universite de Sherbrooke, QC, Canada
[23] Department of Pediatrics, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2060, 3000 CB, Rotterdam, The Netherlands
[24] Institute of Clinical Research, University of Southern Denmark, Odense, Denmark
[25] Centre of Women's, Family and Child Health, University of South-Eastern Norway, Kongsberg, Norway
[26] Department of Biological Sciences, North Carolina State University
[27] Center for Human Health and the Environment, North Carolina State University
[28] Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA USA
[29] Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden
[30] Sachs' Children and Youth Hospital, Södersjukhuset, Stockholm, Sweden
[31] Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, 21205, USA
[32] Department of Woman and Child Health, Ospedale Versilia, Viareggio, Azienda USL Toscana Nord Ovest, Italy
[33] Department of Biochemistry, Université de Sherbrooke, QC, Canada
[34] ECOGENE-21 Biocluster, Chicoutimi Hospital, Saguenay, QC, Canada
[35] Department of Social and Behavioral Science, Harvard TH Chan School of Public Health, 677 Huntington Ave, Boston, MA, USA
[36] Bradford Institute of Health Research, Bradford Royal Infirmary, Bradford BD9 6RJ, UK.
[37] Department of Obstetrics and Gynecology, Duke University Medical Center, Durham, NC 27708
[38] Dept. Translational Research in Psychiatry, Max-Planck-Institute of Psychiatry, Munich, Germany
[39] Centre for Environmental Sciences, Hasselt University
[40] Univ. Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Reproduction and Respiratory Health, IAB, 38000 Grenoble, France
[41] Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, USA
[42] The George Institute for Global Health, Nuffield Department of Women's & Reproductive Health, University of Oxford, Oxford, UK
[43] Grupo de Medicina Xenómica, Fundación Pública Galega de Merdicina Xenómica, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), SERGAS. Santiago de Compostela. Spain
[44] Centro de Investigación en Red de Enfermedades Raras (CIBERER) y Centro Nacional de Genotipado (CEGEN-PRB3). Universidad de Santiago de Compostela. Santiago de Compostela: Spain
[45] Department Public Health &Primary care, Leuven University

[46] Department of Pediatric Oncology and Hematology, Oslo University Hospital, Norway
[47] Norwegian Institute of Public Health, Oslo, Norway
[48] Sachs' Children and Youth Hospital, Södersjukhuset, Stockholm, Sweden
[49] Department of Population Medicine, Harvard Medical School, Harvard Pilgrim Health Care Institute, Boston, MA, USA
[50] Diabetes Unit, Massachusetts General Hospital, Boston, MA, USA
[51] Department of Public Health, Section of Epidemiology, University of Copenhagen, Copenhagen, Denmark
[52] The Novo Nordisk Foundation Center for Basic Metabolic Research, Section on Metabolic Genetics, Faculty of Medical and Health Sciences, University of Copenhagen, Copenhagen, Denmark
[53] Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Research Triangle Park, NC, USA.

*These authors contributed equally to the work

INTRODUCTION

There is growing body of literature demonstrating that the *in utero* environment can impact health later in life.[1-5] DNA methylation is a commonly studied epigenetic mark that can, influence gene expression without change in DNA sequence and is one mechanism through which early life exposures can contribute to the fetal origins of disease.[6] Exposures to chemicals during pregnancy such as tobacco smoke and phthalates, among others, have been associated with differences in umbilical cord blood methylation.[7, 8] Furthermore, site-specific differential methylation of cord blood has also been associated with later-life health outcomes including asthma and insulin sensitivity.[9, 10]

In addition to exposures and health outcomes, inter-individual differences in DNA methylation levels are also impacted by host factors including sex. Prior studies have shown associations of sex with DNA methylation measured in blood at birth (umbilical cord blood),[11-13] in older children,[14, 15] and in adults.[15-19] As expected, there are widespread differences between sexes on the X chromosome CpG sites; however, these studies also reported significant

differences in methylation of autosomes.[13, 15] With the exception of two studies performed primarily in adults[15, 19] and one in children[20], many of the previous studies were limited in sample size with fewer than 200 subjects. It is likely that a much larger meta-analysis focused on umbilical cord DNA methylation would identify many additional sites differentially methylated between boys and girls at birth, a critical developmental period.

In this study, we meta-analyzed cohort-specific associations between sex of the child and Illumina 450K methylation data measured in 8,314 newborn blood samples as well as a follow-up meta-analysis in blood from 5,676 older children within the Pregnancy and Childhood Epigenetics international consortium (PACE). We also investigated enrichment of sex-associated differential methylation in specific biological pathways and diseases. Significant sex differences in disease prevalence, age of onset, and susceptibility across the life course have been observed for various conditions such as asthma, immune response, and metabolic health.[21-23] Therefore, identifying the differences in DNA methylation between boys and girls may highlight the genes that play an active role in the biological mechanisms involved in sex-dependent differences impacting health throughout childhood.

METHODS

Participating cohorts

PACE consists of multiple international birth cohorts with a goal of meta-analyzing exposures and health outcomes in Epigenome Wide Association Studies (EWAS) to understand relationships which may impact maternal and childhood health and disease.[24] Sixteen independent cohorts (N=8,314) contributed data to the analysis of cord blood methylation data and sex, and nine independent cohorts (N=5,676) contributed data (some from multiple time points) to the child methylation analysis. Detailed methods for individual cohorts and

information on which cohorts participated in the cord blood and child analyses are provided in the tables and supplementary methods. All cohorts obtained written informed consent from participants prior to data collection which was approved from local ethics committees.

Methylation Measurement and Quality Control

For all cohorts, DNA methylation was measured using the Illumina Infinium HumanMethylation450 BeadChip.[25] DNA from newborn or child blood samples underwent bisulfite conversion using the EZ-96 DNA Methylation kit (Zymo Research Corporation, Irvine, USA). Methylation quality control and normalization was conducted at the cohort level, as described in the supplementary material. β-values representing proportion of methylation at each CpG site (0 = completely unmethylated, 1 = completely methylated) were used as the methylation outcome. In order to lessen the influence of outlier methylation values, β-values outside a 3IQR range were removed prior to all cohort analyses.

Sex descriptive

Each cohort used recorded child sex, using females as the reference group. As part of quality control, each cohort was asked to check for sex-mismatches using the *getSex* function in the R package *minfi*.[26] This function predicts sex from the median methylation values of probes on the X and Y chromosomes. Any sex mismatches were removed prior to individual cohort analyses. The number of participants for each cohort are reported following the removal of sex mismatches.

Covariates

Cohorts have run two separate models: (1) A crude model adjusting only for batch (and child age in older child models); and (2) the main model, model 1 with additional adjustment for

cell composition. Each study used batch covariates most appropriate for their cohort (e.g. principal components or plate number). Cell composition was estimated using estimateCellCounts in the *minfi* R package.[26] For cord blood analyses, the 'CordBlood' reference data set[27] was used to estimate proportions of 7 cell types (CD8+ T-cells, CD4+ T-cells, NK cells, B-cells, monocytes, granulocytes, and nucleated red blood cells), while older child models used the 'Blood' reference data set[28] which estimates proportions of 6 cell types (CD8+ T-cells, CD4+ T-cells, NK cells, B-cells, monocytes, and granulocytes). Cohorts were also given the option to adjust for genetic ancestry in their models, and this information is included in cohort specific methods. Older child models adjusted for child age at blood draw (years) in all models.

Cohort Specific Statistical Analyses

Each cohort performed independent epigenome-wide association studies (EWAS) according to a common analysis plan approved by all participating cohorts. Models were run using M-type multiple robust linear regression [*rlm()* in the *MASS* R package][29] to control for potential heteroscedasticity and/or influential outliers in the methylation data. In the primary cord blood analysis, the exposure was sex with the outcome of newborn methylation β-values, adjusting for seven estimated cell counts and batch covariates. Two newborn cohorts, NEST and EARLI, ran separate models for subjects of European ancestry and non-European ancestry resulting in two separate datasets for each of these cohorts (N=2 cohorts, N=4 datasets). While most cohorts provided data for the X chromosome, four cohorts were unable to provide this data and were removed from the X chromosome analysis. In the primary older child models, the exposure was sex with the outcome of child methylation beta-values, adjusting for six estimated cell counts, age of the child at blood draw, and batch covariates.

Meta-analysis

All cohorts submitted the results of their cohort level EWAS to the Children's Environmental Health Laboratory (N.Holland-PI) at the University of California, Berkeley. We then performed a fixed effects meta-analysis weighted by the inverse of the variance—using the software METAL[30]—for the main model, which adjusted for seven cell-type proportions and batch. Shadow meta-analyses were conducted independently by co-investigators at the University of Bristol to verify results by L.Kupers. All further analysis was conducted in R version 3.5.2. We excluded SNP control probes (n = 65). The majority of cohorts included probes mapping to the X and Y chromosomes; however, some cohorts were only able to provide results for autosomal probes leaving a total sample size of N = 8,314 for autosomes, and N = 5,213 for subjects with data for sex chromosomes. Filtering of previously identified cross-reactive probes[31] was performed during processing of meta-analysis results. For autosomal probes, this left a total of 456,279 CpG sites measured for association with sex at birth in at least one cohort (331,405 [73%] were measured in all 18 datasets, 423,458 [93%] were measured in at least 17 datasets).

Individual cohorts ran EWAS using R, which in most cases cannot represent numbers smaller than 5e-324. Since many CpG sites were strongly associated with sex, this number was automatically converted to a zero in the R output. To avoid using p-values of zero for the meta-analysis, all zero p-values within a cohort were re-coded as the smallest non-zero p-value for that cohort prior to conducting the meta-analysis.

We adjusted for multiple hypothesis testing using the stringent Bonferroni method, and considered CpG sites with Bonferroni adjusted p-values < 0.05 significant (e.g. 1.1×10^{-7} for 456,279 tests).

Enrichment Analyses

Before enrichment analysis was performed, gene universe was annotated to nearby genes using the *IlluminaHumanMethylation450kanno.ilmn12.hg19* package. Using the differentially methylated CpGs, we performed enrichment analyses at three different levels: pathways, diseases, and molecular signatures. Enrichment for KEGG pathways was done using the *enrichKEGG*() function in the package *clusterProfiler.*[32] Disease enrichment analysis was carried out using *DisGeNET*[33] curated information and the *enricher()* function in *clusterprofiler*.[32] The molecular signatures enrichment included analyses depending on the 15 CpG chromatin states obtained from ROADMAP project[34] and different transcription factors using LOLA.[35] A Bonferroni corrected cutoff of 0.05 was used for significance of pathways, diseases and molecular signatures (e.g. 1.52×10^{-4} for 328 tests for pathways and 4.16×10^{-6} for 12,028 tests for diseases).

RESULTS

Newborns

Results from 16 independent cohorts from the Pregnancy and Childhood Epigenetics (PACE) Consortium, were included in the newborn meta-analysis (N=8,314). Newborn cohort sizes ranged from 53 to 1,319 participants with an average of 462 participants per dataset. There was an even distribution of boys (51%) and girls (49%). Two cohorts, NEST and EARLI, performed separate models for European and non-European participants, resulting in 2 additional datasets (N=18 datasets total). The majority of datasets were made up of participants of European ancestry (N=13 datasets, N=7,067 participants). Other datasets included Hispanic, Mexican-American, African-American, and mixed ethnicities. A summary of the participating newborn cohorts and datasets are included in Table 1.

The results of the individual cohort level newborn models are summarized in Table 2. The average number of sex-associated autosomal CpG sites was 29,303 (after Bonferroni correction), while the average number of sex-associated X chromosome CpG sites was 9,645. For autosome only data, λ ranged from 1.18 to 2.67 with a sample-size weighted average of 1.92.

For the meta-analysis in newborns, there were a total of 46,554 Bonferroni significant sex-associated CpG sites in autosomes out of a total 456,279 total autosomal CpG sites. As expected, the majority of CpG sites on the X chromosome were significantly differentially methylated (N=9,372) between boys and girls. The ten autosomal CpG sites in newborns with the smallest p-values are listed in Table 3. Among these top sites, directionality of methylation changes were in agreement in each individual cohort, suggesting there is no noticeable heterogeneity between the participating studies.

The majority (67%) of sex-associated sites were hypomethylated in boys compared to girls. The Manhattan plot in Figure1a plots shows hypomethylated sites below the null line and hypermethylated sites above the null line. Figure 1a also represents an even distribution of methylation differences throughout the autosomal chromosomes. The CpG-specific difference in methylation level between boys and girls were generally small with a median difference of 0.5% (Figure 2a).

Sites associated with sex in newborns were enriched for many biological processes and diseases. KEGG enrichment analyses showed 59 significantly enriched pathways of 328 tested. KEGG pathways fell into groups containing cancer, signaling, endocrine, addiction, and longevity pathways. Significant KEGG pathways are summarized in Figure 3a with results sorted from most to least significant, and size of circles representing the number of genes included in that pathway. Disease enrichment analyses showed 15 significant diseases of the 12,028 tested. There were

increased CpG sites in genes involved in many disorders, such as mental depression, mood disorders, unipolar disorder, anxiety, substance-abuse, and autism. There were also increased associations for obesity and breast neoplasms (Figure 4a).

Children

For the analysis in older children, data from nine independent cohorts were included (N = 5,995) in the meta-analysis (Table 4). Child cohorts ranged from 124 to 1,053 participants with an average of 516 participants per cohort, and also had an even number of boys (52%) and girls (48%). Similar to the newborn participants, the majority of child datasets contained participants of European ancestry (N=8), with other contributions from Hispanic and Mexican-American cohorts.

The average number of FDR-significant autosomal CpG sites per child cohort was 30,211. For the four older child cohorts that provided X chromosome data, there were an average of 9,345 significant CpG sites. Lambdas for individual autosomal analyses ranged from 1.05 to 5.59, and 1.12 to 5.53 for cohorts with X chromosome data (Table 5).

In older children, there were 46,607 Bonferroni significant autosomal sites associated with child sex. In addition, again the majority of X chromosome sites (8,798) were significant. The top ten sites for older children are listed in Table 6. Similar to newborns, the majority of sites were hypomethylated in older boys compared to older girls which can be seen on the Manhattan plot in Figure 1b. The effect size of differential methylation was small with a median difference of 0.5% among the significant sites in older children (Figure 2b).

KEGG enrichment analyses for pathways in children showed 32 significantly enriched pathways of 328 tested. KEGG pathways predominantly belonged to groups associated with cancer, signaling, and endocrine functions. Significant KEGG pathways are summarized in Figure

3b with results sorted from most to least significant, and size of circles representing the number of genes included in that pathway. Disease enrichment analyses showed 10 diseases of the 12,028 tested to be significant. There were increased CpG sites in genes involved in many mental disorders, such as depression, mood disorders, unipolar disorder, anxiety, memory disorders, and attention deficit hyperactivity disorder (ADHD) (Figure 4b).

Comparison of newborns and older children

The sample size was larger for the newborn analysis compared to the older child analysis (8,314 versus 5,995). There was considerable overlap between significant sites in newborns and children with 70% of child CpG sites also being differentially methylated in newborns (Figure 3). Of these overlapping sites, 99.6% show methylation differences in the same direction indicating that differential methylation is relatively stable over time in children. Similar patterns were seen in both meta-analyses and can also be seen in comparisons of Figures 1a/1b and Figures 2a/2b: 1)a majority were European participants, 2) a large number of significant autosomal CpG sites found, 3)nearly all X chromosome sites significant, 4)the majority of sites were hypomethylated in boys compared to girls, and 5)effect sizes were similarly small and evenly distributed. Pathway and disease enrichment analyses also yielded similar results as can be seen in comparison of figures 3a/3b and 4a/4b. Notably, cancer, signaling, and endocrine pathways dominated both cord blood and child KEGG enrichment analyses, and mental disorders were the most commonly seen in the disease enrichment analyses.

DISCUSSION

This study involving multiple cohorts shows there are widespread differences in methylation of autosomes between boys and girls at birth measured in cord-blood and the majority of these differences persist into later childhood. We report over 40,000 of the nearly 450,000 tested

CpG sites to be differentially methylated with small but statistically significant differences between boys and girls at birth and over 35,000 significant differences in older children with similar effect measurements. In both newborns and children, these differences were enriched in genes involved in cancer pathways and implicated in neurological disorders. This is the first meta-analysis examining differential methylation by sex using the 450K BeadChip. This is also the first analysis of methylation differences between sexes at birth to adjust for cell-type heterogeneity using a cord-blood reference dataset.

We compared our findings to prior studies investigating methylation differences by sex. Only one prior meta-analysis by McCarthy et al.[15] has looked specifically at differential methylation between males and females, and this was assessed using the Illumina 27K chip. Although a few cohorts contributed cord-blood data, the majority of the cohorts included in their analysis used adult blood data. This study reported 184 significant autosomal hits, of which, in our newborn meta-analysis, we replicate 166 (90%) with 95% of hits with methylation change in the same direction. In our child meta-analysis, we replicate 165 (90%) with 94% in the same direction. Another study by Yousefi et al.[13] reported 3,031 CpGs with sex differences in cord-blood for a subset of the CHAMACOS population (which also contributed data to this meta-analysis). Our newborn meta-analysis replicated 2,766 (91%) of the Yousefi et al. hits with 75% of hits in the same direction, and our child meta-analysis replicated 2,723 (90%) with 75% in the same direction. The newborn meta-analysis adds 43,485 autosomal CpGs not previously seen in studies focused on methylation differences by sex with increased sample size and after adjustment of cord blood cell-type heterogeneity. We also report 43,553 new autosomal CpG sites differentially methylated in the blood of older boys and girls.

In both newborns and older children, the majority of the significantly differentially methylated sites were hypomethylated in boys compared to girls, meaning that boys had lower methylation levels than girls. In general, greater gene expression is observed with lower methylation. Hypermethylation in girls is expected in X chromosome CpG sites due to X chromosome inactivation in females; however, these results show that the hypermethylation is not limited to sex-chromosomes. If these methylation changes subsequently impact gene expression, this could mean that genes with methylation differences are being expressed differently in boys compared to girls. These findings agree with the trend previously reported by Yousefi et al. showing hypomethylation in boys for both autosomal and X chromosome sites.

Sex-specific differences are seen in numerous diseases and studies show evidence for a genetic role in sexual dimorphism for disease.[36] Diseases with observed differences by sex include autoimmune diseases,[37] cardiovascular diseases,[38] and pediatric infectious diseases.[23] Early differences between boys and girls also suggests an underlying developmental component.[39] We report many biological pathways and diseases in which the differentially methylated sites are enriched in both newborns and children. Some of the most significantly enriched pathways and diseases have been previously shown to differ between sexes. The top disease pathways included many neurological and mood disorders, and studies have shown that anxiety disorders are more common and more severe in women.[40] Autism is diagnosed in boys more often than girls, and there are differences in the features of autism in each sex.[41] In children, genes involved in ADHD were significantly enriched for differentially methylated CpG sites. ADHD is twice more likely to be diagnosed in boys than girls with different behaviors associated between the sexes.[42] A prior study in the CHAMACOS cohort also reported methylation differences in genes involved in neurological disorders.[13] Our data suggest that DNA methylation may represent one mechanism

contributing to the developmental differences between boys and girls that impact sex-dependent differences in health.

There are several strengths and limitations of our study. We report novel findings of autosomal methylation differences between boys and girls using robust statistical models with a large sample size that was well-powered to assess small effect sizes. We used a new cord blood reference dataset which includes nucleated red blood cells to estimate and adjust for cell-type heterogeneity.[43] All cohorts ensured correct classification of sex prior to analyses using sex chromosome methylation data as a quality control measure. We also included analyses of methylation at two distinct time-points (newborns and older children) which shows the stability of these methylation differences throughout childhood. Although our study included cohorts of multiple ancestries, including European, Hispanic, and African American, the majority of participants were of European ancestry. More work involving a larger number of non-European participants is needed to ensure generalizability of results. Individual cohorts used different normalization methods for methylation data; however, prior studies within the PACE consortium show little difference in final EWAS results from differently normalized data, so we do not expect this to impact the final meta-analysis results.[44] Since this study did not assess if the methylation changes are impacting expression, we cannot confirm if these methylation differences extend to functional changes. These results warrant further follow-up to assess if these methylation changes do indeed impact gene expression in order to confirm the biological significance of these findings and contribution towards the fetal origins of disease hypothesis.[2, 45, 46]

In summary, our study observed many autosomal methylation differences between boys and girls which are enriched in diseases and pathways with differential prevalence between sexes. We replicated and expanded upon previous findings and patterns of autosomal differences and

conducted the largest study to date assessing sex methylation differences in cord blood with additional analysis in child blood. These findings may suggest that early life methylation difference is one potential mechanism through which we see differential disease prevalence.

References

1. Roseboom T, de Rooij S, Painter R. The dutch famine and its long-term consequences for adult health. Early Hum Dev 2006; 82:485-491.

2. Barker DJ. In utero programming of chronic disease. Clin Sci 1998; 95:115-128.

3. Leonard SA, Rasmussen KM, King JC, Abrams B. Trajectories of maternal weight from before pregnancy through postpartum and associations with childhood obesity. Am J Clin Nutr 2017; 106:1295-1301.

4. Kim GH, Berger K, Rauch S, Kogut K, Birgit CH, Antonia MC, Huen K, Eskenazi B, Holland N. Association of prenatal urinary phthalate metabolite concentrations and childhood BMI and obesity. Pediatr Res 2017; 82:405.

5. Slopen N, Loucks EB, Appleton AA, Kawachi I, Kubzansky LD, Non AL, Buka S, Gilman SE. Early origins of inflammation: An examination of prenatal and childhood social adversity in a prospective cohort study. Psychoneuroendocrinology 2015; 51:403-413.

6. Bianco-Miotto T, Craig JM, Gasser YP, van Dijk SJ, Ozanne SE. Epigenetics and DOHaD: From basics to birth and beyond. J Dev Orig Health Dis 2017; 8:513-519.

7. Solomon O, Yousefi P, Huen K, Gunier RB, Escudero-Fung M, Barcellos LF, Eskenazi B, Holland N. Prenatal phthalate exposure and altered patterns of DNA methylation in cord blood. Environ Mol Mutagen 2017; 58:398-410.

8. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu C, Küpers LK, Oh SS, Hoyo C, Gruzieva O, Söderhäll C, Salas LA, Baïz N, Zhang H, Lepeule J, Ruiz C, Ligthart S, Wang T, Taylor JA, Duijts L, Sharp GC, Jankipersadsing SA, Nilsen RM, Vaez A, Fallin MD, Hu D, Litonjua AA, Fuemmeler BF, Huen K, Kere J, Kull I, Munthe-Kaas M, Gehring U, Bustamante M, Saurel-Coubizolles M, Quraishi BM, Ren J, Tost J, Gonzalez JR, Peters MJ, Håberg SE, Xu Z, van Meurs JB, Gaunt TR, Kerkhof M, Corpeleijn E, Feinberg AP, Eng C, Baccarelli AA, Benjamin Neelon SE, Bradman A, Merid SK, Bergström A, Herceg Z, Hernandez-Vargas H, Brunekreef B, Pinart M, Heude B, Ewart S, Yao J, Lemonnier N, Franco OH, Wu MC, Hofman A, McArdle W, Van dV, Falahi F, Gillman MW, Barcellos LF, Kumar A, Wickman M, Guerra S, Charles M, Holloway J, Auffray C, Tiemeier HW, Smith GD, Postma D, Hivert M, Eskenazi B, Vrijheid M, Arshad H, Antó JM, Dehghan A, Karmaus W, Annesi-Maesano I, Sunyer J, Ghantous A, Pershagen G, Holland N, Murphy SK, DeMeo DL, Burchard EG, Ladd-Acosta C, Snieder H, Nystad W, Koppelman GH, Relton CL, Jaddoe VWV, Wilcox A, Melén E, London SJ. DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. Am J Hum Genet 2016; 98:680-696.

9. den Dekker HT, Burrows K, Felix JF, Salas LA, Nedeljkovic I, Yao J, Rifas-Shiman S, Ruiz-Arenas C, Amin N, Bustamante M, DeMeo DL, Henderson AJ, Howe CG, Hivert MF, Ikram MA, de Jongste JC, Lahousse L, Mandaviya PR, van Meurs JB, Pinart M, Sharp GC, Stolk L, Uitterlinden AG, Anto JM, Litonjua AA, Breton CV, Brusselle GG, Sunyer J, Smith GD, Relton CL, Jaddoe VWV, Duijts L. Newborn DNA-methylation, childhood lung function, and the risks of asthma and COPD across the life course. Eur Respir J 2019; 53:10.1183/13993003.017-2018. Print 2019 Apr.

10. van Dijk SJ, Peters TJ, Buckley M, Zhou J, Jones PA, Gibson RA, Makrides M, Muhlhausler BS, Molloy PL. DNA methylation in blood from neonatal screening cards and the association with BMI and insulin sensitivity in early childhood. Int J Obes (Lond) 2018; 42:28-35.

11. Adkins RM, Thomas F, Tylavsky FA, Krushkal J. Parental ages and levels of DNA methylation in the newborn are correlated. BMC medical genetics 2011; 12:47.

12. Adkins RM, Krushkal J, Tylavsky FA, Thomas F. Racial differences in gene-specific DNA methylation levels are present at birth. Birth Defects Res A Clin Mol Teratol 2011; 91:728-736.

13. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. BMC Genomics 2015; 16:911.

14. Fuke C, Shimabukuro M, Petronis A, Sugimoto J, Oda T, Miura K, Miyazaki T, Ogura C, Okazaki Y, Jinno Y. Age related changes in 5-methylcytosine content in human peripheral leukocytes and placentas: An HPLC-based study. Ann Hum Genet 2004; 68:196-204.

15. Nina SM, Phillip EM, Cadby G, Yazar S, Franchina M, Eric KM, David AM, Alex WH. Meta-analysis of human methylation data for evidence of sex-specific autosomal patterns. BMC Genomics 2014; 15:981.

16. Inoshita M, Numata S, Tajima A, Kinoshita M, Umehara H, Yamamori H, Hashimoto R, Imoto I, Ohmori T. Sex differences of leukocytes DNA methylation adjusted for estimated cellular proportions. Biol Sex Differ 2015; 6:1-7. eCollection 2015.

17. Liu J, Morgan M, Hutchison K, Vince DC. A study of the influence of sex on genome wide methylation. PLoS One 2010; 5:e10028.

18. Marco PB, Eske MD, Daniel JW, Strengman E, Janson E, Iris ES, René SK, Roel AO. The relationship of DNA methylation with age, gender and genotype in twins and healthy controls. PLoS One 2009; 4:e6767.

19. Singmann P, Shem-Tov D, Wahl S, Grallert H, Fiorito G, Shin SY, Schramm K, Wolf P, Kunze S, Baran Y, Guarrera S, Vineis P, Krogh V, Panico S, Tumino R, Kretschmer A, Gieger C, Peters A, Prokisch H, Relton CL, Matullo G, Illig T, Waldenberger M, Halperin E. Characterization of whole-genome autosomal differences of DNA methylation between men and women. Epigenetics Chromatin 2015; 8:4-3. eCollection 2015.

20. Suderman M, Simpkin A, Sharp G, Gaunt T, Lyttleton O, McArdle W, Ring S, Davey Smith G, Relton C. Sex-associated autosomal DNA methylation differences are wide-spread and stable throughout childhood. bioRxiv 2017:118265.

21. Dearden L, Bouret SG, Ozanne SE. Sex and gender differences in developmental programming of metabolism. Mol Metab 2018; 15:8-19.

22. Arathimos R, Granell R, Henderson J, Relton CL, Tilling K. Sex discordance in asthma and wheeze prevalence in two longitudinal cohorts. PLoS One 2017; 12:e0176293.

23. Muenchhoff M, Goulder PJR. Sex differences in pediatric infectious diseases. J Infect Dis 2014; 209 Suppl 3:120.

24. Felix JF, Joubert BR, Baccarelli AA, Sharp GC, Almqvist C, Annesi-Maesano I, Arshad H, Baïz N, Bakermans-Kranenburg M, Bakulski KM, Binder EB, Bouchard L, Breton CV, Brunekreef B, Brunst KJ, Burchard EG, Bustamante M, Chatzi L, Cheng Munthe-Kaas M, Corpeleijn E, Czamara D, Dabelea D, Davey Smith G, De Boever P, Duijts L, Dwyer T, Eng C, Eskenazi B, Everson TM, Falahi F, Fallin MD, Farchi S, Fernandez MF, Gao L, Gaunt TR, Ghantous A, Gillman MW, Gonseth S, Grote V, Gruzieva O, Håberg SE, Herceg Z, Hivert M, Holland N, Holloway JW, Hoyo C, Hu D, Huang R, Huen K, Järvelin M, Jima DD, Just AC, Karagas MR, Karlsson R, Karmaus W, Kechris KJ, Kere J, Kogevinas M, Koletzko B, Koppelman GH, Küpers LK, Ladd-Acosta C, Lahti J, Lambrechts N, Langie SAS, Lie RT, Liu AH, Magnus MC, Magnus P, Maguire RL, Marsit CJ, McArdle W, Melén E, Melton P, Murphy SK, Nawrot TS, Nisticò L, Nohr EA, Nordlund B, Nystad W, Oh SS, Oken E, Page CM, Perron P, Pershagen G, Pizzi C, Plusquin M, Raikkonen K, Reese SE, Reischl E, Richiardi L, Ring S, Roy RP, Rzehak P, Schoeters G, Schwartz DA, Sebert S, Snieder H, Sørensen TIA, Starling AP, Sunyer J, Taylor JA, Tiemeier H, Ullemar V, Vafeiadi M, Van Ijzendoorn MH, Vonk JM, Vriens A, Vrijheid M, Wang P, Wiemels JL, Wilcox AJ, Wright RJ, Xu C, Xu Z, Yang IV, Yousefi P, Zhang H, Zhang W, Zhao S, Agha G, Relton CL, Jaddoe VWV, London SJ. Cohort profile: Pregnancy and childhood epigenetics (PACE) consortium. Int J Epidemiol 2017.

25. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL, Fan J, Shen R. High density DNA methylation array with single CpG site resolution. Genomics 2011; 98:288-295.

26. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: A flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 2014; 30:1363-1369.

27. Andrews SV BK. FlowSorted.CordBlood.450k: Illumina 450k data on sorted cord blood cells. 2019; R package version 1.12.0.

28. Jaffe AE. FlowSorted.blood.450k: Illumina HumanMethylation data on sorted blood cell populations. 2019; R package version 1.22.0.

29. Venables WN, Ripley BD. *Modern Applied Statistics with S.* 4th ed. New York: Springer-Verlag; 2002.

30. Huber W, Carey VJ, Gentleman R, Anders S, Carlson M, Carvalho BS, Bravo HC, Davis S, Gatto L, Girke T, Gottardo R, Hahne F, Hansen KD, Irizarry RA, Lawrence M, Love MI, MacDonald J, Obenchain V, Oleś A,K., Pagès H, Reyes A, Shannon P, Smyth GK, Tenenbaum D, Waldron L, Morgan M. Orchestrating high-throughput genomic analysis with bioconductor. Nature methods 2015; 12:115-121.

31. Chen Y, Lemire M, Choufani S, Butcher DT, Grafodatskaya D, Zanke BW, Gallinger S, Hudson TJ, Weksberg R. Discovery of cross-reactive probes and polymorphic CpGs in the illumina infinium HumanMethylation450 microarray. Epigenetics 2013; 8:203-209.

32. Yu G, Wang L, Han Y, He Q. clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology 2012; 16:284-287.

33. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res 2017; 45:D83-D839.

34. Roadmap EC, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y, Pfenning A, Wang X, ClaussnitzerYaping Liu M, Coarfa C, Alan Harris R, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, David Hawkins R, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Scott Hansen R, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K, Feizi S, Karlic R, Kim A, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong N, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Abdennur N, Adli M, Akerman M, Barrera L, Antosiewicz-Bourget J, Ballinger T, Barnes MJ, Bates D, Bell RJA, Bennett DA, Bianco K, Bock C, Boyle P, Brinchmann J, Caballero-Campo P, Camahort R, Carrasco-Alfonso M, Charnecki T, Chen H, Chen Z, Cheng JB, Cho S, Chu A, Chung W, Cowan C, Athena Deng Q, Deshpande V, Diegel M, Ding B, Durham T, Echipare L, Edsall L, Flowers D, Genbacev-Krtolica O, Gifford C, Gillespie S, Giste E, Glass IA, Gnirke A, Gormley M, Gu H, Gu J, Hafler DA, Hangauer MJ, Hariharan M, Hatan M, Haugen E, He Y, Heimfeld S, Herlofsen S, Hou Z, Humbert R, Issner R, Jackson AR, Jia H, Jiang P, Johnson AK, Kadlecek T, Kamoh B, Kapidzic M, Kent J, Kim A, Kleinewietfeld M, Klugman S, Krishnan J, Kuan S, Kutyavin T, Lee A, Lee K, Li J, Li N, Li Y, Ligon KL, Lin S, Lin Y, Liu J, Liu Y, Luckey CJ, Ma YP, Maire C, Marson A, Mattick JS, Mayo M, McMaster M, Metsky H, Mikkelsen T, Miller D, Miri M, Mukame E, Nagarajan RP, Neri F, Nery J, Nguyen T, O'Geen H, Paithankar S, Papayannopoulou T, Pelizzola M, Plettner P, Propson NE, Raghuraman S, Raney BJ, Raubitschek A, Reynolds AP, Richards H, Riehle K, Rinaudo P, Robinson JF, Rockweiler NB, Rosen E, Rynes E, Schein J, Sears R, Sejnowski T, Shafer A, Shen L, Shoemaker R, Sigaroudinia M, Slukvin I, Stehling-Sun S, Stewart R, Subramanian SL, Suknuntha K, Swanson S, Tian S, Tilden H, Tsai L, Urich M, Vaughn I, Vierstra J, Vong S, Wagner U, Wang H, Wang T, Wang Y, Weiss A, Whitton H, Wildberg A,

Witt H, Won K, Xie M, Xing X, Xu I, Xuan Z, Ye Z, Yen C, Yu P, Zhang X, Zhang X, Zhao J, Zhou Y, Zhu J, Zhu Y, Ziegler S, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, Amin V, Whitaker JW, Schultz MD, Ward LD, Sarkar A, Quon G, Sandstrom RS, Eaton ML, Wu Y, Pfenning AR, Wang X, Claussnitzer M, Liu Y, Coarfa C, Harris RA, Shoresh N, Epstein CB, Gjoneska E, Leung D, Xie W, Hawkins RD, Lister R, Hong C, Gascard P, Mungall AJ, Moore R, Chuah E, Tam A, Canfield TK, Hansen RS, Kaul R, Sabo PJ, Bansal MS, Carles A, Dixon JR, Farh K, Feizi S, Karlic R, Kim A, Kulkarni A, Li D, Lowdon R, Elliott G, Mercer TR, Neph SJ, Onuchic V, Polak P, Rajagopal N, Ray P, Sallari RC, Siebenthall KT, Sinnott-Armstrong N, Stevens M, Thurman RE, Wu J, Zhang B, Zhou X, Beaudet AE, Boyer LA, De Jager PL, Farnham PJ, Fisher SJ, Haussler D, Jones SJM, Li W, Marra MA, McManus MT, Sunyaev S, Thomson JA, Tlsty TD, Tsai L, Wang W, Waterland RA, Zhang MQ, Chadwick LH, Bernstein BE, Costello JF, Ecker JR, Hirst M, Meissner A, Milosavljevic A, Ren B, Stamatoyannopoulos JA, Wang T, Kellis M. Integrative analysis of 111 reference human epigenomes. Nature 2015; 518:317.

35. Sheffield NC, Bock C. LOLA: Enrichment analysis for genomic region sets and regulatory elements in R and bioconductor. Bioinformatics 2016; 32; 2015/10/27:587-589.

36. Ober C, Gilad Y, Loisel DA. Sex-specific genetic architecture of human disease. Nature Reviews Genetics 2008; 9:911-922.

37. Whitacre CC. Sex differences in autoimmune disease. Nat Immunol 2001; 2:777-780.

38. Kander MC, Cui Y, Liu Z. Gender difference in oxidative stress: A new look at the mechanisms for cardiovascular diseases. J Cell Mol Med 2017; 21:1024-1032.

39. Uekert SJ, Akan G, Evans MD, Li Z, Roberg K, Tisler C, Dasilva D, Anderson E, Gangnon R, Allen DB, Gern JE, Lemanske RF. Sex-related differences in immune development and the expression of atopy in early childhood. J Allergy Clin Immunol 2006; 118:1375-1381.

40. McLean, Carmen P.|Asnaani, Anu|Litz, Brett T.|Hofmann,Stefan G. Gender differences in anxiety disorders: Prevalence, course of illness, comorbidity and burden of illness. J Psychiatr Res 2011; 45:1027-1035.

41. Halladay AK, Bishop S, Constantino JN, Daniels AM, Koenig K, Palmer K, Messinger D, Pelphrey K, Sanders SJ, Singer AT, Taylor JL, Szatmari P. Sex and gender differences in autism spectrum disorder: Summarizing evidence gaps and identifying emerging areas of priority. Molecular autism 2015; 6:36.

42. Bauermeister JJ, Shrout PE, Chávez L, Rubio-Stipec M, Ramírez R, Padilla L, Anderson A, García P, Canino G. ADHD and gender: Are risks and sequela of ADHD the same for boys and girls? Journal of Child Psychology and Psychiatry 2007; 48:831-839.

43. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, L McKenney S, Witter F, Walston J, Feinberg AP, Fallin MD. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics 2016; 11:354-362.

44. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, Reese SE, Markunas CA, Richmond RC, Xu C, Küpers LK, Oh SS, Hoyo C, Gruzieva O, Söderhäll C, Salas LA, Baïz N, Zhang H, Lepeule J, Ruiz C, Ligthart S, Wang T, Taylor JA, Duijts L, Sharp GC, Jankipersadsing SA, Nilsen RM, Vaez A, Fallin MD, Hu D, Litonjua AA, Fuemmeler BF, Huen K, Kere J, Kull I, Munthe-Kaas M, Gehring U, Bustamante M, Saurel-Coubizolles M, Quraishi BM, Ren J, Tost J, Gonzalez JR, Peters MJ, Håberg SE, Xu Z, van Meurs JB, Gaunt TR, Kerkhof M, Corpeleijn E, Feinberg AP, Eng C, Baccarelli AA, Benjamin Neelon SE, Bradman A, Merid SK, Bergström A, Herceg Z, Hernandez-Vargas H, Brunekreef B, Pinart M, Heude B, Ewart S, Yao J, Lemonnier N, Franco OH, Wu MC, Hofman A, McArdle W, Van dV, Falahi F, Gillman MW, Barcellos LF, Kumar A, Wickman M, Guerra S, Charles M, Holloway J, Auffray C, Tiemeier HW, Smith GD, Postma D, Hivert M, Eskenazi B, Vrijheid M, Arshad H, Antó JM, Dehghan A, Karmaus W, Annesi-Maesano I, Sunyer J, Ghantous A, Pershagen G, Holland N, Murphy SK, DeMeo DL, Burchard EG, Ladd-Acosta C, Snieder H, Nystad W, Koppelman GH, Relton CL, Jaddoe VWV, Wilcox A, Melén E, London SJ. DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. Am J Hum Genet 2016; 98:680-696.

45. Essex MJ, Armstrong JM, Thomas Boyce W, Hertzman C, Lam LL, Sarah MAN, Kobor MS. Epigenetic vestiges of early developmental adversity: Childhood stress exposure and DNA methylation in adolescence. Child Dev 2013; 84:58-75.

46. Armstrong DA, Lesseur C, Conradt E, Lester BM, Marsit CJ. Global and gene-specific DNA methylation across multiple tissues in early infancy: Implications for children's health research. FASEB J 2014; 28:2088-2097.

Metabolism Clinical and Experimental 110 (2020) 154292

Contents lists available at ScienceDirect

Metabolism Clinical and Experimental

journal homepage: www.metabolismjournal.com

Translational

A multi-omic analysis of birthweight in newborn cord blood reveals new underlying mechanisms related to cholesterol metabolism

Rossella Alfano [a,b,c,1], Marc Chadeau-Hyam [a,b,d,1], Akram Ghantous [e,2], Pekka Keski-Rahkonen [e,2], Leda Chatzi [f,g], Almudena Espin Perez [h], Zdenko Herceg [e], Manolis Kogevinas [i,j,k,l], Theo M. de Kok [m], Tim S. Nawrot [c,n], Alexei Novoloaca [e], Chirag J. Patel [o], Costanza Pizzi [p], Nivonirina Robinot [e], Franca Rusconi [q], Augustin Scalbert [e], Jordi Sunyer [j,k,l], Roel Vermeulen [a,d], Martine Vrijheid [i,j,l], Paolo Vineis [a,b,r], Oliver Robinson [a,3], Michelle Plusquin [a,b,c,3,*]

[a] Department of Epidemiology and Biostatistics, The School of Public Health, Imperial College London, London, United Kingdom
[b] Medical Research Council-Health Protection Agency Centre for Environment and Health, Imperial College London, London, United Kingdom
[c] Centre for Environmental Sciences, Hasselt University, Diepenbeek, Belgium
[d] Institute for Risk Assessment Sciences (IRAS), Division of Environmental Epidemiology, Utrecht University, Utrecht, the Netherlands
[e] International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, 69008 Lyon, France
[f] Department of Preventive Medicine, University of Southern California, Los Angeles, CA 90007, United States
[g] Department of Social Medicine, University of Crete, Heraklion, Crete, Greece
[h] Department of Biomedical Informatics Research, Stanford University, CA, United States
[i] Consortium for Biomedical Research in Epidemiology and Public Health (CIBERESP), Madrid, Spain
[j] ISGlobal, Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
[k] Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain
[l] Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[m] Department of Toxicogenomics, Maastricht University, Maastricht, the Netherlands
[n] Environment & Health Unit, Leuven University, Leuven, Belgium
[o] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States
[p] Department of Medical Sciences, University of Turin and CPO-Piemonte, Torino, Italy
[q] Unit of Epidemiology, Anna Meyer Children's University Hospital, Florence, Italy
[r] Human Genetic Foundation (HuGeF), Turin, Italy

ARTICLE INFO

Article history:
Received 18 November 2019
Accepted 11 June 2020

Keywords:
Birth weight
Cholesterol
DNA methylation
Gene expression
Metabolome
Proteins

ABSTRACT

Background: Birthweight reflects in utero exposures and later health evolution. Despite existing studies employing high-dimensional molecular measurements, the understanding of underlying mechanisms of birthweight remains limited.
Methods: To investigate the systems biology of birthweight, we cross-sectionally integrated the methylome, the transcriptome, the metabolome and a set of inflammatory proteins measured in cord blood samples, collected from four birth-cohorts ($n = 489$). We focused on two sets of 68 metabolites and 903 CpGs previously related to birthweight and investigated the correlation structures existing between these two sets and all other omic features via bipartite Pearson correlations.
Results: This dataset revealed that the set of metabolome and methylome signatures of birthweight have seven signals in common, including three metabolites [PC(34:2), plasmalogen PC(36:4)/PC(O-36:5), and a compound with *m/z* of 781.0545], two CpGs (on the *DHCR24* and *SC4MOL* gene), and two proteins (periostin and CCL22). CCL22, a macrophage-derived chemokine has not been previously identified in relation to birthweight. Since the results of the omics integration indicated the central role of cholesterol metabolism, we explored the association of cholesterol levels in cord blood with birthweight in the ENVIRONAGE cohort ($n = 1097$), finding that higher birthweight was associated with increased high-density lipoprotein cholesterol and that high-density lipoprotein cholesterol was lower in small versus large for gestational age newborns.

Abbreviations: AGA, adequate for gestational age; BMI, body mass index; DOHaD, Developmental Origin of Health and Disease; HDL, high-density lipoprotein; IL, interleukin; IQR, interquartile; LGA, large for gestational age; LDL, low-density lipoprotein; *m/z*, mass-to-charge ratio; ORA, overrepresentation analysis; SGA, small for gestational age; PC, phosphatidylcholine; U, unassigned metabolite; 95CI, 95% confidence interval.

* Corresponding author at: Centre for Environmental Sciences, Hasselt University, Agoralaan, building D, 3590 Diepenbeek, Belgium.
E-mail address: michelle.plusquin@uhasselt.be (M. Plusquin).

[1]Joint first authors.
[2]Joint second authors.
[3]Joint last authors.

https://doi.org/10.1016/j.metabol.2020.154292

Conclusions: Our data suggests that an integration of different omic-layers in addition to single omics studies is a useful approach to generate new hypotheses regarding biological mechanisms. CCL22 and cholesterol metabolism in cord blood play a mechanistic role in birthweight.

1. Introduction

The Developmental Origin of Health and Disease hypothesis (DOHaD) states that later life diseases may be influenced by experiences and conditions in prenatal life [1]. It is hypothesized that the interplay between genotype and in utero environmental factors induces molecular modifications and possibly phenotype differentiation in the fetus [2]. This developmental plasticity is of crucial importance for postnatal life but can induce impairments related to adverse health outcomes [3]. For example, exposure to detrimental environmental factors may induce birthweight changes that in turn may be associated with increased mortality or risk of cardiovascular diseases, mental health problems, and some cancers later in life [4–7]. Among several mechanisms proposed to explain the link between in utero exposures, birthweight and health and diseases in later life, molecular markers identified through "omics" platforms, may play a central role. The theoretical foundation that drives the present study is that birthweight induces molecular modifications that in turn influence later life health, as exemplified in the Fig. 1, however is also possible that birthweight is influenced by molecular changes determined by in utero exposures.

Recently two studies based on high-dimensional molecular measurements in cord blood identified DNA methylation signals and metabolites associated with birthweight: 914 differentially methylated CpG sites were discovered in 8825 neonates from 24 birth-cohorts and 68 metabolites were identified in 499 neonates from four birth-cohorts [8,9]. These studies, together with several metabolomic, gene expression, proteomic, genomic, and epigenomic analyses, have increased our understanding of underlying mechanisms of birthweight [10–20]. However, a study integrating multi-omic levels in cord blood associated with birthweight in the same samples has not yet been performed.

To decipher at several levels the molecular cascades that regulate birthweight, in this paper we propose to integrate DNA-methylation, gene expression data as well as metabolic profiles and a set of inflammatory proteins measured in cord blood samples ($n = 489$) collected from four independent population-based birth-cohorts [21]. We used two birthweight-related sets of molecular signals, metabolites [8] and DNA methylation levels [9], to drive the integrated analyses and translated these signals to the other omic layers in the same samples. Based on results common to the metabolite- and methylation-driven multi-omic integrations, we aimed to identify key molecular associations with birthweight.

2. Material and methods

2.1. Study population and samples collection

Our study population arises from the EXPOsOMICS European project and includes 500 newborns from four population-based cohorts: 200 newborns from ENVIRONAGE, 100 from INMA, 99 from Piccolipiù, and 101 from Rhea [21–25]. Inclusion criteria and protocols are detailed in the respective cohort descriptions and in the Supplementary methods. Before the placenta was delivered, whole blood was withdrawn from cord vessels and immediately frozen at −80 °C. Samples were sent to different laboratories for metabolome, inflammatory proteins and DNA methylome analysis [21]. The transcriptome was measured for the 200 ENVIRONAGE samples participating in the EXPOsOMICS project and cholesterol was measured for the entire ENVIRONAGE cohort.

2.2. Metabolomic profiles

Untargeted metabolomics was performed as previously described [8] and detailed in the Supplementary Methods. Briefly, reversed phase liquid chromatography-quadrupole time-of-flight mass spectrometry (UHPLC-QTOF-MS) system was used in positive ion mode with 499 of the 500 samples successfully analyzed. Raw data preprocessing was performed with Agilent MassHunter software, and metabolic features present in <60% of the samples were excluded, leaving 4712 features for 499 samples available for the subsequent analysis. Data were log-transformed and missing values were imputed using the impute. QRILC function within the "imputeLCMD" R package. Identification of the features of interest was done as previously described by Robinson et al. [8] and level of identification was reported as proposed by Sumner et al. [26].

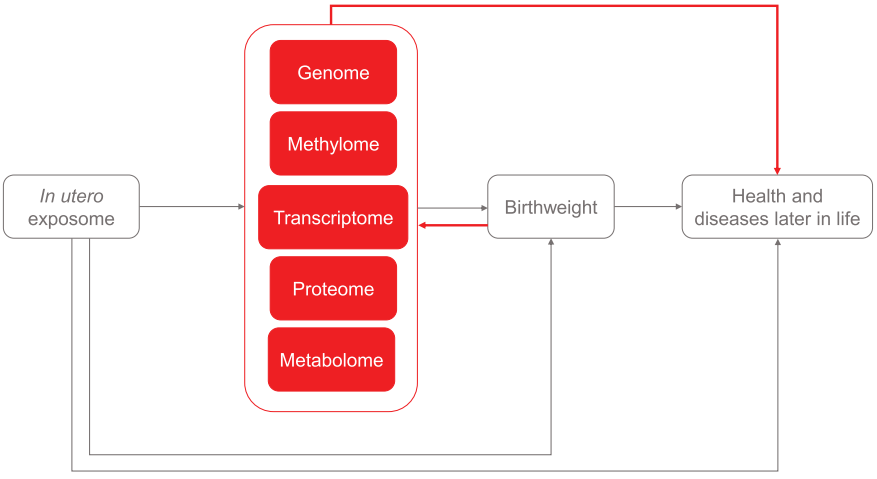


Fig. 1. Research hypothesis. The diagram exemplifies the central role of omics in the hypothetical path that drives the present study and according to which birthweight induces molecular modifications that in turn influence the later life health and disease (red arrows); and the alternative pathways where in utero exposures cause the changes in the omic layers leading to variation of birthweight (grey arrows).

2.3. DNA methylation profiles

Genomic DNA was extracted from buffy coats according to standard protocol and underwent bisulphite conversion using the Zymo EZ DNA methylationTM kit (Zymo, Irvine, CA, USA), hybridization to Illumina HumanMethylation 450K BeadChip arrays and scanning using an Illumina iScan. As detailed in the Supplementary Methods and elsewhere, we used in-house software to preprocess the data, including removal of probes based on signal intensities and control probes, background subtraction and dye bias correction [27]. After quality control and filtering, 417572 CpG sites for 460 samples were retained for subsequent analysis and methylation levels were expressed as beta values. To account for technically-induced and tissue-related variation in the methylation levels, we ran a preliminary linear model for the methylation beta values (as outcome variable) adjusting for technical variation (array row and position on the chip), as well as estimated cell type composition using the Bakulski method [28].

2.4. Gene expression

Gene expression levels were measured in cord blood samples ($n = 200$) of the ENVIRONAGE cohort. RNA was extracted using the total RNA miRNeasy mini kit (Qiagen, Venlo, Netherlands) according to the manufacturer's protocol. As detailed in the Supplementary Methods, samples were quality checked and further hybridized onto Agilent Whole Human Genome 8 × 60 K microarrays coupled with Agilent DNA G2505C Microarray Scanner. After preprocessing, quality control, and normalization detailed in the Supplementary methods, 29164 transcripts for 165 samples were left available for further analysis. To account for technical noise, we ran a linear model for the observed gene expression level (as outcome variable) adjusting for technical variation (hybridization date) and white blood cells count.

2.5. Inflammatory proteins

We measured 22 inflammation-related proteins using an R & D Luminex screening assay according to the protocol described by the manufacturer, and c-reactive protein using Solid Phase Sandwich ELISA. Excluding proteins detected in <40% of the study population, we were left with 16 inflammatory proteins (Supplementary Table 1) measured in 493 samples, of which three were excluded. For the remaining 490 samples missing values were imputed following an approach based on likelihood maximization estimation procedure [29]. Protein concentrations were subsequently log-transformed, and to correct for nuisance variation, we employed the same linear model approach described before setting the plate identification as technical covariate.

2.6. Cholesterol

The plasma levels of high-density lipoprotein (HDL), low-density lipoprotein (LDL), and total cholesterol were measured in entire ENVIRONAGE population using Cobas 8000 C702 module analyzer (Roche, Basel, Switzerland). Outliers (>5 standard deviations from mean) were excluded from the analyses. Respectively 1139, 1109 and 1131 samples had valid HDL, LDL, and total cholesterol measurements.

2.7. Anthropometrics and covariates

Birthweight in grams was collected from medical records. In ENVIRONAGE cohort, newborns were classified as small for gestational age (SGA), adequate for gestational age (AGA) or large for gestational age (LGA) if their birthweight for given gestational age, sex, and parity status was respectively below the 10th percentile, between 10th percentile and 90th percentile, or above the 90th percentile calculated for Flanders from the Study Centre for Perinatal Epidemiology (http://www.neonatologie.ugent.be/SPE-standaarden.pdf).

As detailed in the Supplementary Material, covariates were selected based on previous reported associations with birthweight and included: sex of the newborns, parity, gestational age, maternal and paternal age and body mass index (BMI), maternal smoking status during pregnancy, and maternal education.

2.8. Statistical analyses

The study design is schematically represented in Fig. 2.

2.8.1. Exploring correlation structure across omic profiles

We adopted an exposome globe approach to investigate the correlation structures across the omic measurements available in our study population and investigated Pearson's correlation coefficients for pairs of omics measurement [30]. We used sets of molecular features from two omic platforms to drive our integrated analyses [31]: (a) a set of 68 metabolites previously associated with birthweight (Supplementary Table 2); and (b) a set of 903 (available from the total of 914, Supplementary Table 2) CpGs previously associated with birthweight [8,9]. These sets were correlated with all the other untargeted omic features. The statistical significance of all correlation coefficients was assessed by deriving a z-score from Fisher transformation and running a Student's *t*-test test assessing the null hypothesis of no correlation $H_0 : \rho = 0$. We corrected for multiple testing using the stringent Bonferroni correction for the number of tests (i.e. the total number of pairs investigated) and considered significant the correlations with Bonferroni corrected p-values<0·05. The number of samples participating in each analysis is presented in Supplementary Table 3. Results were visualized by means of Circos plots (Circos software version 0.69–6), where only significant correlation coefficients were reported.

In order to assess if our results were biased by heterogeneity between the different cohorts, we used linear models adjusted on the factors differing across cohorts (namely: gestational age, parental ages, weights and heights, parity and maternal education) to test associations between features significantly correlated in our main analyses. Stratification by sex was performed as a sensitivity analysis in order to take into account the birthweight sexual dysmorphism [32].

Finally, we compared the metabolite- and the methylation-driven results to assess if any metabolite-CpG pairs and any omic were in common.

2.8.2. Pathway analysis

We performed overrepresentation analyses (ORA) of all transcripts and CpGs significantly correlated in the metabolite-driven analyses and of the CpG significantly correlated with metabolites in the metabolite-driven analysis using ConsensuspathDB online tool (http://consensuspathdb.org/). A pathway was considered significantly enriched if p-values were smaller than 0·05 and included at least 3 genes.

Enriched metabolic pathways within metabolic features correlated with CpGs in the methylation-driven analyses were identified using the *mummichog* program (version 1.1.0) [33] through the MetaboAnalyst platform [34]. We used all mass-to-charge ratio (m/z) values and associated p-values of the metabolic features as software input and set *mummichog* parameters to 'positive mode' at ±5 ppm mass tolerance. The p-value cutoff to identify the list of significant m/z features was set to false discovery rate adjusted (FDR) p-value equal to 0·05, with the non-FDR significant features used as the reference set. The algorithm searches tentative compound lists from metabolite reference databases against an integrated model of human metabolism to identify functional activity. A pathway was considered significant if gamma adjusted p-values were smaller than 0·05. Visualization of enriched pathways on the KEGGscape network was performed through the MetaboAnalyst platform [34].

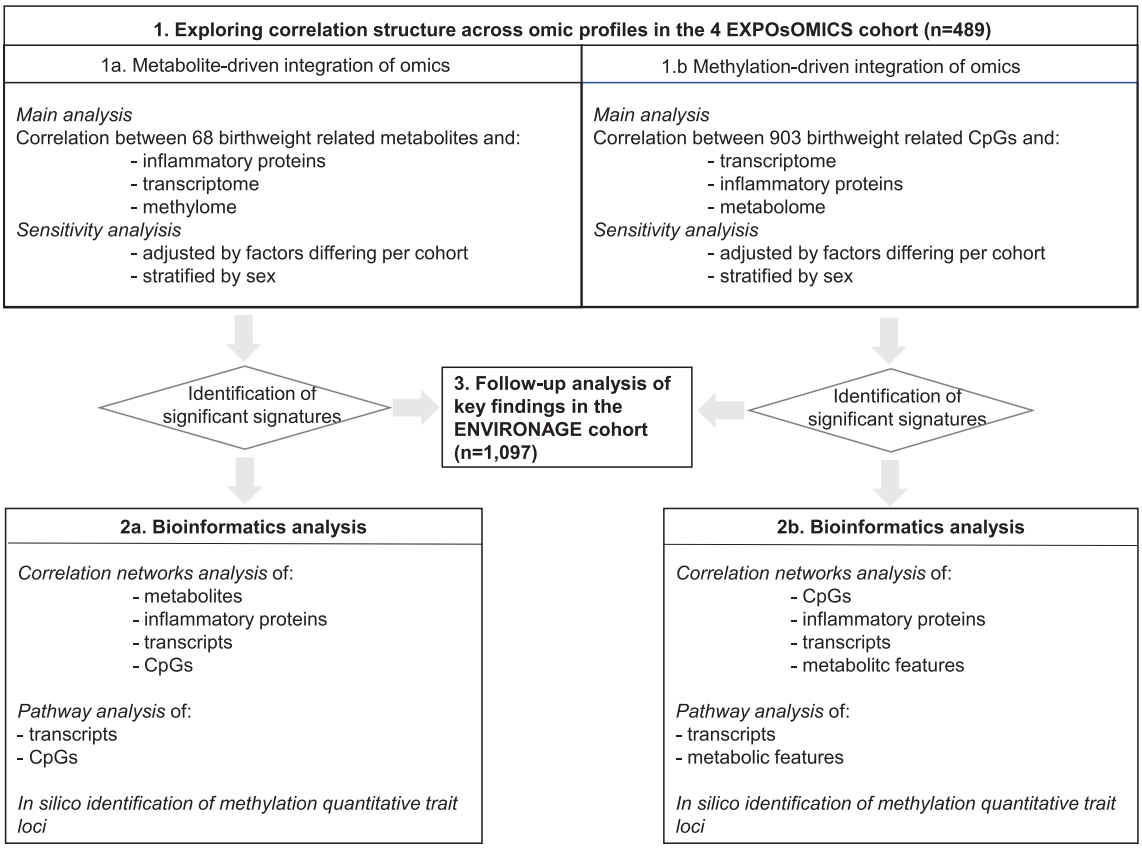


Fig. 2. Study design. The figure shows the main analysis exploring the structure across omic profiles in the four EXPOsOMICS cohorts and the subsequent pathways analysis, network correlation analysis, in silico identification of methylation quantitative trait loci, and the follow-up analysis exploring key findings in the ENVIRONAGE cohort.

2.8.3. Correlation network analysis

We selected all significantly correlated omic features in the main analyses and performed network correlation analysis. Correlation networks were plotted using "igraph" package with layout algorithm by Fruchterman and Reingold (version 0.7.1). Only nodes with degree >2 and edges with Bonferroni corrected p-values <0·05 were represented in the networks. Communities were detected using the Louvain algorithm.

2.8.4. In silico identification of methylation quantitative trait loci

We searched SNPs associated with the CpG sites significantly correlated to the other omic layers in the publicly available methylation quantitative trait loci (mQTL) database (http://www.mqtldb.org/). We compared the localization on the genome of 233 SNPs associated with own birthweight and 128 SNPs associated with offspring birthweight in the NHGRI-EBI GWAS Catalog with the identified methylation signals (±2 Mb from the genetic variants) and we searched for mQTL associated with these SNPs [35].

2.8.5. Exploring key findings in the ENVIRONAGE cohort

Based on omics identified in significant pairs from both the metabolite- and methylation-driven analyses, we generated a hypothesis on mechanisms underlying birthweight. Common omic signals that were significantly correlated with both the metabolite and CpGs birthweight-related sets, were associated with cholesterol metabolism. We ran linear regression models to assess (i) the associations between omics identified in our multi-omic analyses and cholesterol levels and (ii) the association between birthweight and cholesterol levels. These analyses were restricted to the ENVIRONAGE cohort. All the models were adjusted for gestational age, parity, newborn sex, maternal age, maternal height, maternal BMI, smoking during pregnancy and maternal education, total cholesterol levels (for HDL and LDL analyses), plate (for proteins analyses), cell types composition, array row and position on the chip (for CpGs analyses). Paternal age and anthropometric measurements were not included as adjustment covariates due to missingness. In addition, we used linear models adjusted for the covariates aforementioned to explore if being SGA, AGA and LGA (independent variable) was associated with the levels of omic markers identified and cholesterol (dependent variable). If p-values were smaller than 0·05 results were considered significant. Samples participating in each analysis are reported in the Supplementary Table 3.

3. Results

3.1. Population

Descriptive characteristics of the study population participating in the main multi-omic study are presented by cohort in Table 1 and by sex in Supplementary Table 4 and indicate heterogeneity across cohorts for all covariates, except proportion of girls born and maternal smoking habits during pregnancy.

3.2. Metabolite-driven integration of omics

By correlating the set of ($n = 68$) metabolites (Supplementary Table 2), previously reported in these same four birth-cohorts to be associated with birthweight by Robinson and colleagues, with the other omic layers we identified 347 significantly correlated omic pairs involving 208 omic features (47 metabolites, 15 inflammatory proteins, 71 transcripts and 75 CpGs) (Fig. 3A and Supplementary Table 5) [8].

Pairs involving transcripts were all positively correlated to metabolites (Supplementary Figs. 1A and 2B), 68% of the metabolite-

Table 1
Characteristics of the EXPOsOMICS population (by cohort) and the full ENVIRONAGE population.

	EXPOsOMICS population n = 489				P-value[a]	Full ENVIRONAGE population n = 1097	P-value[b]
	ENVIRONAGE *n* = 195	INMA *n* = 97	Piccolipiù n = 97	RHEA *n* = 100			
Birthweight, g	3389·28 ± 478·29	3305·98 ± 399·33	3217·06 ± 431·42	3258·90 ± 429·69	**8·85e−03**	3413·01 ± 468·77	**4·645e−05**
Birthweight,					–		–
SGA (<10thPi)	14 (7·2)	–	–	–		76 (6·9)	
AGA (≥10thPi& ≤ 90thPi)	155 (79·5)	–	–	–		880 (80·2)	
LGA (>90th Pi)	25 (12·8)	–	–	–		141 (12·8)	
Gestational age, weeks	39·14 ± 1·53	39·71 ± 1·41	39·57 ± 1·58	38·43 ± 1·32	**2·66e−09**	39·22 ± 1·48	0·70
Girls	95 (49·0)	50 (51·5)	43 (44·3)	47 (47·0)	0·77	530 (48·3)	0·99
Maternal age, years	29·41 ± 4·43	31·48 ± 4·14	33·28 ± 4·46	30·03 ± 4·99	**8·25e−11**	29·41 ± 4·58	**4·05e−07**
Maternal BMI, Kg/m^2	23·94 ± 4·06	23·45 ± 3·83	22·63 ± 3·87	25·09 ± 5·37	**7·87e−04**	24·52 ± 4·78	**3·90e−03**
Maternal weight, Kg	66·09 ± 11·88	62·52 ± 11·20	60·95 ± 11·16	66·76 ± 15·64	**9·28e−04**	67·82 ± 14·26	**3·75e−06**
Maternal height, cm	166·14 ± 6·81	163·15 ± 6·60	164·05 ± 5·70	162·93 ± 5·65	**3·22e−05**	166·189 ± 6·48	**1·52e−06**
Maternal smoking	25 (12·9)	23 (24·0)	20 (20·6)	20 (20·2)	0·09	134 (12·2)	**2·40e−03**
Maternal education					**1·57e−03**		**0·02**
Primary school	27 (14·6)	17 (17·5)	8 (8·2)	8 (8·1)		139 (12·7)	
Secondary school	63 (34·1)	46 (47·4)	40 (41·2)	57 (57·6)		393 (35·8)	
University of higher	95 (51·4)	34 (35·1)	49 (50·5)	34 (34·3)		565 (51·5)	
Multiparity	87 (44·8)	43 (44·8)	51 (52·6)	70 (71·4)	**1·16e−04**	511 (46·5)	0.06
Paternal age, years	31·75 ± 5·89	33·46 ± 4·35	36·47 ± 5·57	34·24 ± 5·04	**1·94e−10**	31·95 ± 5·47	**4·86e−07**
Paternal BMI, Kg/m^2	25·78 ± 3·46	27·20 ± 3·90	24·97 ± 3·03	25·99 ± 4·49	**1·06e−03**	25·96 ± 3·87	0·88
Paternal weight, Kg	83·43 ± 15·94	81·11 ± 13·36	78·43 ± 10·65	84·97 ± 14·47	**5·97e−03**	83·65 ± 13·72	0·09
Paternal height, cm	179·07 ± 7·54	177·08 ± 6·80	177·20 ± 6·30	176·38 ± 7·21	**0·01**	179·49 ± 7·33	**8·08e−05**

Counts (percentages) and means ± standard deviations are reported for categorical and continuous variables, respectively.
AGA = adequate for gestational age; LGA = large for gestational age; Pi = percentile calculated for Flanders from the Study Centre for Perinatal Epidemiology; SGA = small for gestational age.
[a] P-value for associations between the four EXPOsOMICs birth-cohort. P-values <0·05 are marked in bold.
[b] Between the pooled EXPOsOMICs population and the full ENVIRONAGE population are detected with analysis of variance test (for continuous variables) and chi square test (for categorical variables). P-values <0·05 are marked in bold.

inflammatory protein pairs were positively correlated (Supplementary Figs. 1C and 2C) while 82% of the significant metabolite-CpG pairs showed negative correlation coefficients (Supplementary Figs. 1B and 2A). The strongest correlation coefficients in absolute value were observed in pairs involving transcripts (absolute range r = 0·42–0·54), followed by pairs involving CpG sites (absolute range r = 0·28–0·39), and inflammatory proteins (absolute range r = 0·18–0·44) (Table 2).

We did not identify one single metabolite related to all three of the other omic layers (Fig. 3B and Supplementary Table 6). Progesterone was the annotated metabolite correlated with most omic features (n = 47), including both proteins and transcripts.

The identified transcriptome signals were involved in 31 significant pathways (p-values<0·05), mainly related to immune response [e.g. Interleukin (IL)12-mediated signaling events and natural killer cell mediated cytotoxicity] (Fig. 3C and Supplementary Table 7). Similarly, the identified CpGs signals were mapped into seven significant pathways (p-values<0·05) including TNFalfa, thermogenesis, and insulin signaling pathways (Fig. 3C and Supplementary Table 7).

To further characterize the 208 omic signals identified in the metabolite-driven analysis, we constructed a correlation network. This network analyses revealed that identified molecules were mainly grouped in distinct communities according to their omic layer (Fig. 3D). In all groups, except group 4, signals of other omics were also present, e.g. progesterone and other four unassigned metabolites (U4, U5, U6 and U46) that were grouped with the transcripts (group 2) (Supplementary Table 8). The network analysis unveiled novel correlations between proteins and CpGs groups but not between transcripts and CpGs or transcripts and proteins.

No mQTL was identified for the identified 75 CpG sites. One to six significant CpG sites were located ±2 Mb from 153 SNPs out of the 233 associated with own birthweight and from 39 SNPs out of the 128 SNPs associated with offspring birthweight (Supplementary Table 9). One SNP, associated with own birthweight, was located in the same gene (*PDE4B*) as a significant CpG.

Sensitivity analyses agreed mostly with the main analyses, except for transcripts. 29 metabolites were still significantly associated with 32 omic features (17 CpG sites and 15 proteins) after adjustment for gestational age, parental ages, parental weights and heights, parity and maternal education (Supplementary Table 10). Out of the total 347 significant correlations from the main analyses, after stratification by sex, 62 correlations remained significant in boys and 46 in girls (Supplementary Fig. 3), three additional correlations became significant in boys and five in girls (Supplementary Table 11).

3.3. Methylation-driven integration of omics

By correlating a set of 903 CpG sites that have been previously associated with birthweight by Kupers and colleagues (Supplementary Table 2) with the other omic layers, we identified 482 significant pairs involving 241 omic measurements (58 CpGs, 157 transcripts, two proteins, 24 metabolic features) (Fig. 4A, Supplementary Table 12) [9].

As indicated in Table 3, most of the CpG-transcript and CpG-inflammatory protein pairs were negatively correlated (92% and 75% respectively, see Supplementary Figs. 4A,C and 5A–B), and the pairs involving metabolites were predominantly positively correlated (79%, Supplementary Figs. 4B and 5C).

The strongest correlations were observed in the CpG-transcript pairs (absolute range r = 0·45–0·57), followed by CpG-metabolic feature pairs (absolute range r = 0·26–0·39) and CpG-inflammatory protein pairs (absolute range r = 0·22–0·24).

No CpG site was correlated to all three types of omic data (Fig. 4D and Supplementary Table 13). *cg08217545* (located on the *NFIC* gene) was the CpGs involved in the most significant correlations (n = 86). Pathways analysis of the identified transcriptome signals resulted in 17 enriched pathways (Fig. 4B and Supplementary Table 14), mostly involved in signal transduction and immune system (such as G beta: gamma signaling through PI3Kgamma and TNF signaling pathway).

The 24 metabolic features identified were found to represent 14 unique compounds, of which nine could be identified [Unidentifiable phosphatidylcholine (PC)/LysoPC, PC(30:0), PC(34:2), PC(36:4), Plasmalogen PC(38:4) or PC(O-38:5), Plasmalogen PC(36:4) or PC(O-36:5), Plasmalogen PC(36:3) or PC(O-36:4), Cholesterol, Cholestenone].

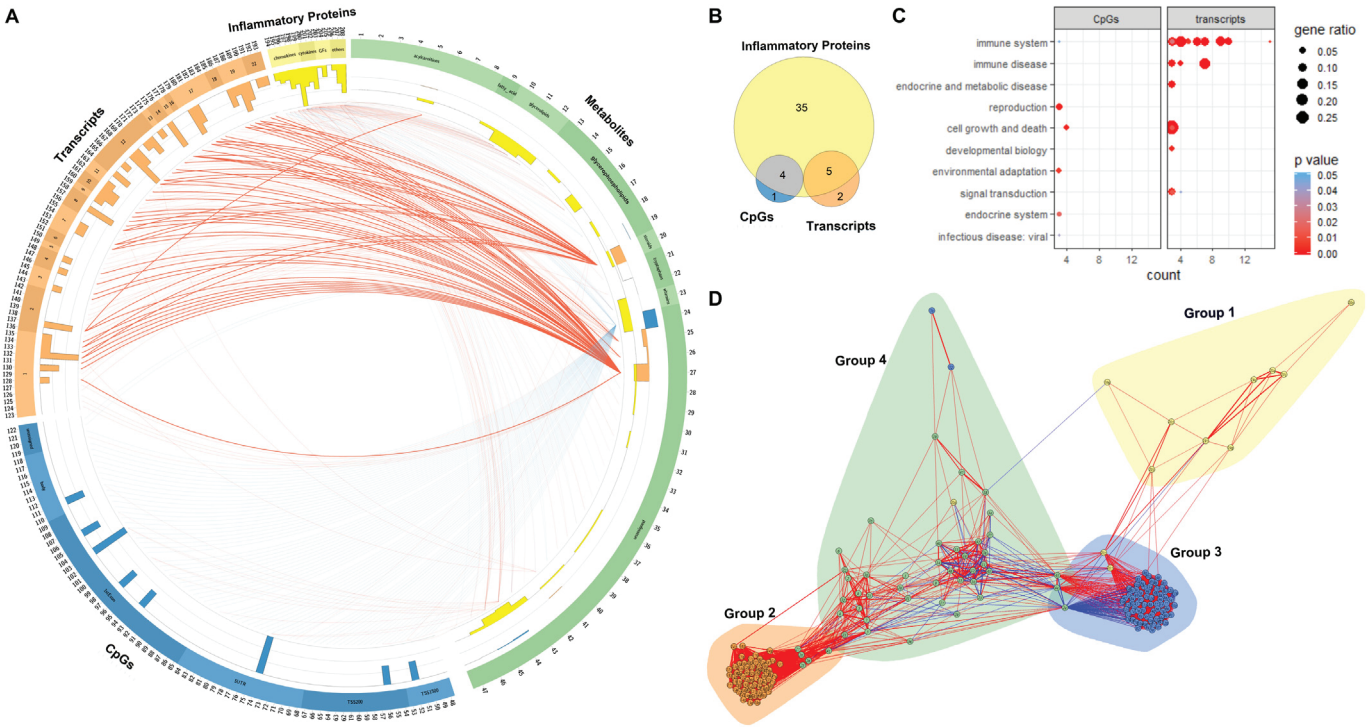


Fig. 3. Significant correlations in the metabolite-driven integration of omics. A The circular plot displays the results of metabolite-driven integration of omics. Tracks from outside to inside are: ideogram, histogram plot and significant links between omic signals. The ideogram shows the omic features significantly correlated grouped in metabolites (green), CpGs (blue), transcripts (orange) and inflammatory proteins (yellow). On the ideogram, features are identified by numbers as reported in the Supplementary Table 27. Alternating bands distinguish classes of metabolites and proteins, genomic regions of genes associated to CpGs and chromosomes on which are located the genes associated to transcripts. The histogram shows for each omic feature (on the x-axis) the scaled percentage of significant correlations per each omic set as identified by colors (on the y-axis in increasing order from outside to inside). In the center of the circular plot each significant correlation coefficient is visualized through a link connecting the two correlated omics. Links are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Thickness of the links grows according to increasing absolute value of correlation coefficients. B The venn diagram shows the count of metabolites significant in the metabolite-inflammatory protein, metabolite-transcriptome and metabolite-methylome analyses represented in yellow, orange and blue circles respectively, and their intersections. C The dot plot shows significant pathways grouped by function, from overrepresentation analysis (ORA) of the transcripts and the CpGs identified in the metabolite-driven integration of omics. Dots size varies according the gene ratio and colors according the p-values. D The network chart shows results from correlation network analysis. The nodes represent the omic features. Nodes are colored according the omic layer they belong to (metabolites in green, CpGs in blue, transcripts in orange and inflammatory proteins in yellow) and are identified by numbers as reported in the Supplementary Table 27. Only nodes with degree>2 and edges with Bonferroni corrected p-values <0·05 are displayed. Hedges are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Communities (groups 1–4), detected using the Louvain algorithm, are marked by circles. TSS = transcription start site; UTR = untranslated region; GFs = growth factors.

Full details on retention times masses, and levels of identification are reported in Supplementary Table 15, with the chromatograms and mass spectra in the Additional data. Pathways analysis for the metabolic signals identified ($n = 201$ metabolic features with 0·05 FDR adjusted p-values) revealed three significantly enriched pathways including C21-steroid hormone biosynthesis and metabolism, porphyrin metabolism and omega-3 fatty acid metabolism (Fig. 4C).

Network correlation analysis identified omics grouped in three multi-omic communities of transcripts and CpGs (groups 1, 2 and 5) and three communities mainly or uniquely populated by a single omic type, e.g. groups 3 and 4 are only made of transcripts (Fig. 4E and Supplementary Table 16). The network analysis unveiled novel correlations between metabolites and proteins, and metabolites and transcripts.

Three mQTLs were identified for the 58 significant CpG sites (Supplementary Table 17). None of these three mQTLs has been previously associated with birthweight. We identified one to seven significant CpG sites located ±2 Mb from 188 SNPs out of the 233 associated with own birthweight and from 96 SNPs out of the 128 associated with offspring birthweight (Supplementary Table 18). Only one SNP associated with own birthweight was located in the same gene (*PIM3*) as a significant CpG site.

Table 2
Overview of the metabolite-driven integration of omics.

	Metabolite-inflammatory protein pairs	Metabolite-transcript pairs	Metabolite-CpG site pairs
Samples	489	164	460
Correlation pairs	1088	1,983,152	28,394,896
P-value threshold	4·60e−05	2·52e−08	1·76e−09
	Significant correlations		
Correlation pairs	133 (12·2)	129 (0·006)	85 (0·0002)
Omic features	44 metabolites (64·71) 15 proteins (93·75)	7 metabolites (10·29) 71 transcripts (0·24)	5 metabolites (7·35) 75 CpGs (0·02)
r absolute values range	0·18–0·44	0·42–0·54	0·28–0·39
Negative r	43 (32·33)	0 (0)	70 (82·35)

Counts (percentages) are reported. r = correlation coefficient.

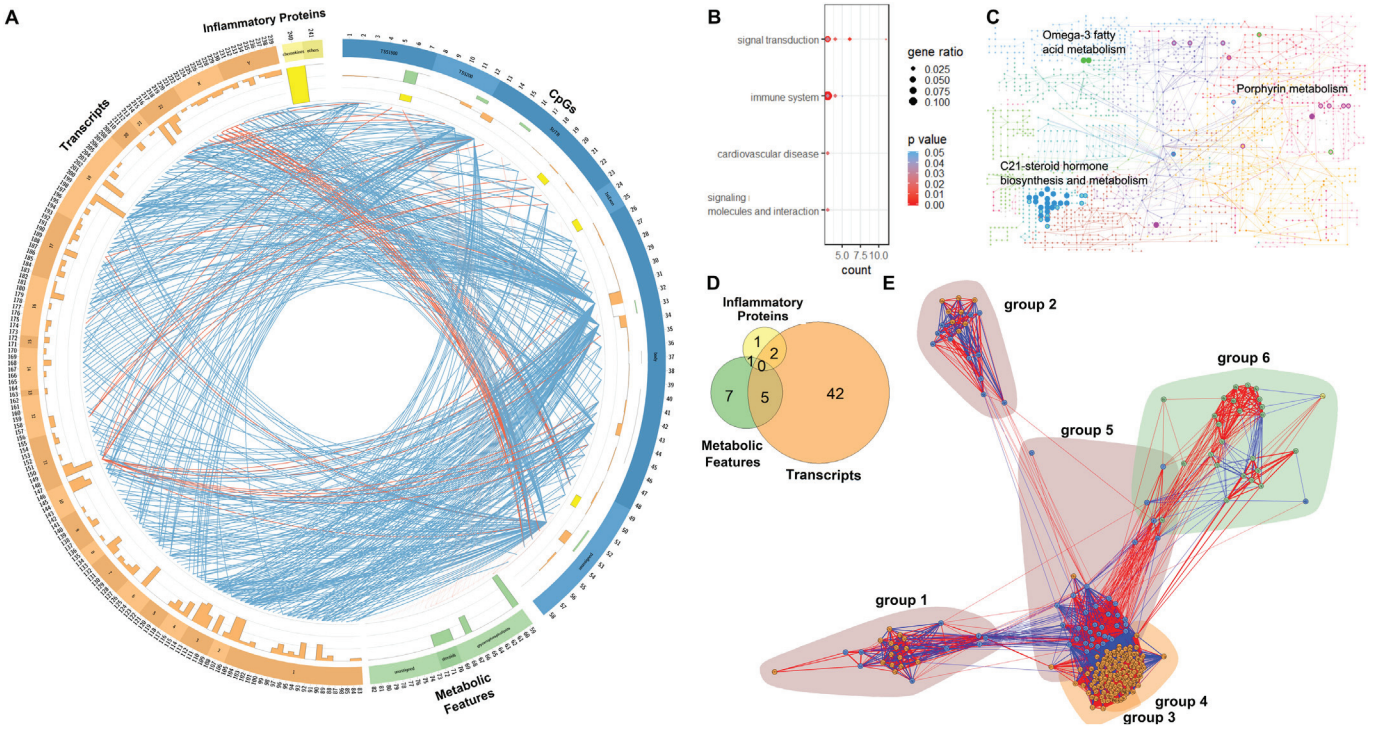


Fig. 4. Significant correlations in the methylation-driven integration of omics. A The circular plot depicts the results of the methylation-driven integration of omics. Tracks from outside to inside are: ideogram, histogram plot and significant links between omic signals. The ideogram shows the omic features significantly correlated grouped in CpGs (blue), metabolic features (green), transcripts (orange) and inflammatory proteins (yellow). On the ideogram, features are identified with numbers as reported in the Supplementary Table 28. Alternating bands distinguish classes of metabolites and proteins, genomic regions of genes associated to CpGs and chromosomes on which are located the genes associated to transcripts. The histogram shows for each omic feature (on the x-axis) the scaled percentage of significant correlations per each omic set as identified by colors (on the y-axis in increasing order from outside to inside). In the center of the circular plot each significant correlation coefficient is visualized through a link connecting the two correlated omics. Links are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Thickness of the links grows according to increasing absolute value of correlation coefficients. B The dot plot shows significant pathways, grouped by function, from overrepresentation analysis (ORA) of the transcripts identified in the methylation-driven integration of omics. Dots size varies according the gene ratio and colors according the p-values. C Metabolic network visualization of significantly enriched pathways based on the manually curated KEGG global metabolic network [34]. The metabolites of significantly enriched pathways are represented as nodes on the network. Empty nodes represent compounds identified from the feature list by *mummichog* but not significant, while solid nodes represent significantly enriched features. Note not all metabolites from the KEGG global network are displayed. D The venn diagram shows the count of significant CpGs in the CpG-inflammatory protein, CpG -transcriptome and CpG-metabolic feature analyses represented in yellow, orange and green circles respectively, and their intersections. E The network chart shows results from correlation network analysis. The nodes represent the omic features. Nodes are colored according the omic layer they belong to (metabolites in green, CpGs in blue, transcripts in orange and inflammatory proteins in yellow) and are identified by numbers as reported in the Supplementary Table 28. Only nodes with degree>2 and edges with Bonferroni corrected p-values <0·05 are displayed. Hedges are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Communities (groups 1–6), detected using the Louvain algorithm, are marked by circles. TSS = transcription start site; UTR = untranslated region.

In sensitivity analyses 13 CpGs were still significantly associated with 24 omics (seven metabolic features and 17 transcripts) after adjustment for gestational age, parental ages, weights and heights, parity and maternal education (Supplementary Table 19). After stratification by sex, of the 482 correlations significant in the main analyses four remained significant in boys and nine in girls (Supplementary Fig. 6), four additional correlations became significant in boys and four in girls (Supplementary Table 20).

3.4. Signals in common between the metabolite- and methylation-driven integration of omics

We identified seven features in common to both the metabolite- and the methylation-driven integration of omics (Fig. 5A and Supplementary Table 21) which are part of 48 unique correlation pairs.

The seven features include three metabolites [PC(34:2), plasmalogen PC(36:4)/PC(O-36:5), and an unidentifiable compound

Table 3
Overview of the methylation-driven integration of omics.

	Methylation-inflammatory protein pairs	Methylation-transcript pairs	Methylation-metabolic feature pairs
Samples	450	162	460
Correlation pairs	14,448	26,335,092	4,254,936
P-value threshold	3·46e−06	1·89e−09	1·17−08
		Significant correlations	
Correlation pairs	4 (0·03)	439 (0·002)	39 (0·0009)
Omic features	4 CpGs (0·44)	49 CpGs (5·43)	13 CpGs (1·44)
	2 proteins (12·5)	157 transcripts (0·54)	24 metabolic features (0·51)
r absolute values range	0·22–0·24	0·45–0·57	0·26–0·39
Negative r	3 (75)	403 (91·80)	8 (20·51)

Counts (percentages) are reported. r = correlation coefficient.

(U)61 of *m/z* 781.0545], two CpG sites (*cg17901584* on the *DHCR24* gene, and *cg05119988* on the *SC4MOL* gene), and two proteins (CCL22 and periostin).

No feature was in common between all omic-layers (Fig. 5B).

Despite that no single transcript was in common, pathways related to immune system and signal transduction were enriched in both metabolite- and the methylation-driven integration of omics (Fig. 5D) and the "chemokine signaling pathway" in particular was significant in both analyses, along with pathways involving IL-2 and JAK-STAT signaling (Supplementary Tables 7 and 14).

Network correlation analysis confirmed the three metabolites were correlated with the two CpGs (Fig. 5C).

No feature was found to be robust to adjustment for factors differing by cohorts (Supplementary Table 22).

As both genes in which the common CpG sites are located are involved in cholesterol biosynthesis we hypothesized that cholesterol metabolism is associated with birthweight and the newly identified omics signals and consequently followed-up our analyses with a verification study in the ENVIRONAGE cohort.

3.5. Cholesterol analysis

To test our hypothesis that cholesterol is related to birthweight and birthweight related molecules, we analyzed the measured levels of each of the seven omics features, common to both the metabolite- and the methylation-driven omic integration, in relation to cord blood measurements of cholesterol available in the ENVIRONAGE cohort (Supplementary Tables 3 and 23). Regressions models were adjusted for gestational age, parity, newborn sex, maternal age, maternal height, maternal BMI, smoking during pregnancy, maternal education, total cholesterol levels (for HDL and LDL analyses), plate (for proteins analyses), cell types composition, array row and position on the chip (for CpGs analyses). We found that an interquartile (IQR) increment of all three metabolic features [PC(34:2), plasmalogen PC(36:4)/PC(O-36:5) and U61], and the two CpG sites, *cg05119988* (on the *SC4MOL* gene) and *cg17901584* (on the *DHCR24* gene), were respectively associated with an increase in total cholesterol levels of 17, 27, 18, 6, 7 mg/dl (p-value<0·01 for all the associations) (Table 4). Additionally, a IQR-increment of plasmalogen PC(36:4)/PC(O-36:5) was associated with an increase of 5 mg/dl of HDL cholesterol (p-value = 0·01) (Table 4).

Analyses stratified by sex showed similar results for total cholesterol in girls and boys (only the p-value of association with CpGs lost statistical significance in girls). Associations involving HDL cholesterol lost statistical significance in the girls' analyses (Supplementary Table 24).

In a larger subset of the ENVIRONAGE cohort (n = 1096) (Supplementary Tables 3 and 23), we found that an IQR-increment in birthweight (equal to 618 g) was associated with an increment 1·14 mg/dl of HDL cholesterol [95% confidence interval (95CI) = 0·43 mg/

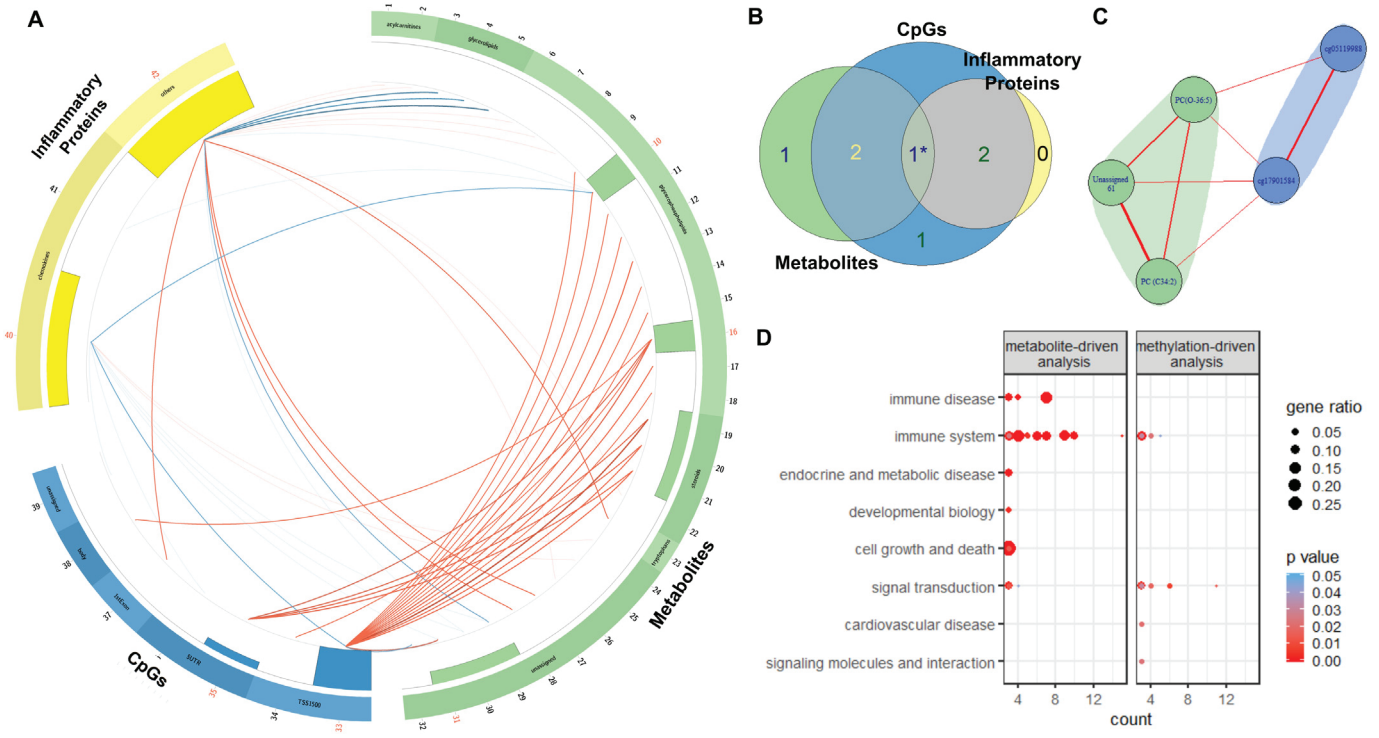


Fig. 5. Significant correlations common to the metabolite- and the methylation-driven integrations of omics. The figure represents the 48 significant correlations identified by the seven omics commonly significant in the metabolite- and the methylation-driven integrations of omics. A In the circular plot tracks from outside to inside are: ideogram, histogram plot and significant links between omic signals. The ideogram shows the omic features significantly correlated grouped in metabolites (green), CpGs (blue) and inflammatory proteins (yellow). On the ideogram features are identified by numbers as reported in the Supplementary Table 29. Alternating bands distinguish classes of metabolites and proteins, and genomic regions of genes associated to CpGs. The seven common omic features are highlighted in red. The histogram plot shows for each omics feature (on the x-axis) the scaled percentage of significant correlations (on the y-axis in increasing order from outside to inside). In the center of the circular plot each significant correlation coefficient is visualized through a link connecting the two correlated omics. Links are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Thickness of the links grows according to increasing absolute value of correlation coefficients. B The venn diagram shows among the seven common omic features how many are significantly correlated with inflammatory proteins, transcripts and metabolites represented in yellow, orange and green circles respectively, and their intersections. C The network chart shows results from correlation network analysis. The nodes represent the omic features. Nodes are colored according the omic layer they belong to (metabolites in green, and CpGs in blue) and are identified by omic names. Only nodes with degree>2 and edges with Bonferroni corrected p-values <0·05 are displayed. Hedges are colored according to the sign of the correlation coefficient, where red and blue links mean respectively positive and negative correlations. Communities (groups 1–2), detected using the Louvain algorithm, are marked by circles. D The dot plot shows significant pathways grouped by function, from overrepresentation analysis (ORA) of the transcripts identified in the metabolite- and methylation-driven integration of omics. Dots size varies according the gene ratio and colors according the p-values. TSS = transcription start site; UTR = untranslated region * intersection only between proteins and metabolites.

dl to 1·85 mg/dl, p-value = 1·71e-03] upon adjusted for gestational age, parity, newborn sex, maternal age, maternal height, maternal BMI, smoking during pregnancy and maternal education (Supplementary Table 25). Analyses stratified by sex confirmed this association in girls only (Supplementary Table 25).

Finally, we tested if the levels of omic markers identified and cholesterol differ in SGA, LGA and AGA newborns. Among the seven omic markers identified only methylation of the two CpG sites, *cg05119988* (on the *SC4MOL* gene) and *cg17901584* (on the *DHCR24* gene), was significantly higher in SGA compared to LGA (p-value = 0·03 and 0·01, respectively), and methylation of *cg17901584* was significantly lower in LGA compared to AGA (p-value = 0·03) (Supplementary Table 26). Total cholesterol was significantly lower (estimate change = −4·68 mg/dl, 95CI = −8·84 mg/dl to −0·50 mg/dl) in SGA compared to AGA newborns (p-value = 0·03) and HDL cholesterol was significantly lower (estimate change = −2·17 mg/dl, 95CI = −4·24 mg/dl to −0·09 mg/dl) in SGA compared to LGA newborns (p-value = 0·04) (Fig. 6). In the analyses stratified by sex, HDL cholesterol levels were lower in SGA girls compared to both AGA (p-value = 0·05) and LGA (p-value = 0·03) (Supplementary Table 25).

4. Discussion

Through an in-depth exploration of birthweight-associated sets of metabolites and methylation sites we have identified commonalities and differences between signals from two different molecular layers [8,9]. From millions of correlations between omic molecules measured in cord blood, our study shows that the set of metabolome and methylome signatures of birthweight have seven signals in common, one of which, the macrophage-derived chemokine CCL22, has not been previously identified in relation to birthweight. CCL22 was negatively correlated to both the metabolite PC(C34:2) and *cg17901584* on the *DHCR24* gene. CCL22 plays a crucial role in the control of T cell immunity [36]. Similarly, we found that the "Chemokine signaling pathway" identified through the gene expression analysis, overlaps between the metabolite- and methylation-driven analyses, supporting a potential link between birthweight and the immune system.

Although a detailed discussion of specific molecules is beyond the purpose of the present study and requires further experimental validation, we highlight here the example of progesterone, which was the annotated metabolite most frequently correlated in metabolite-driven analyses. Higher levels of progesterone in cord blood are observed with lower-weight births [8,37]. Progesterone plays an important role in the suppression of immune responses promoting cord blood T cell differentiation [38]. Furthermore, we observed that progesterone was clustered with transcripts that were most enriched for the IL-12 signaling pathway, which promotes Th1 differentiation and forms a link between innate resistance and adaptive immunity [39]. In the methylation-driven analyses, JAK3 was the most frequently correlated transcript and the JAK-STAT signaling pathway, which plays a critical role orchestrating innate and adaptive immunity, was also significant enriched in the gene expression pathway analysis [40]. No previous study has linked this enzyme to birthweight, yet experimental evidence has associated JAK3 with low grade inflammation, obesity and metabolic syndrome [41]. Progesterone and JAK3 are two signals of many we identified, that illustrate how the cross-omic approach can provide deeper insight in the biological underlying mechanisms of birthweight and highlight avenues for further investigation.

In general, by comparing the metabolite- and methylation-driven analyses we could observe that: i) phosphatidylcholine metabolites, particularly plasmalogens, have been identified in both analyses. Maternal plasmalogens, that are able to cross the placenta, have been recently associated with newborn body composition [42]. However, most previous studies identified lysoPCs rather than PCs as dominant cord blood metabolites related to birthweight [15,18]. ii) In both analyses, metabolites were grouped with two CpGs, albeit different CpGs (*cg05119988* and *cg17901584* in the metabolite-driven analysis and *cg14195992* and *cg15331996* in the methylation-driven analysis). iii) The metabolite-driven network analysis showed relations to groups of features from distinct omic layers while in the methylation-driven analysis three groups contain a mix of different omic types, representing a more profound multi-omics signal. iv) In both analyses we consistently found stronger correlations with gene expression-methylation and gene expression-metabolites than between the other layers. v) Both analyses identified different gene expression signals which suggests that post-transcriptional regulation is specific for metabolites or methylation signals. vi) Metabolite-CpG pairs in the metabolite-driven analyses were mainly negatively correlated, while in the methylation-driven analyses the opposite was observed. The latter difference may arise from the fact that expression levels of genes are positively correlated with the level of methylation within the transcribed region and while only 12% of the CpGs related to metabolites in the metabolite-driven analysis were located in the gene body - this percentage increased to 46% in the methylation-driven analysis. In the methylation-driven analyses the various negative associations may be due to the large proportion of negative correlations between transcripts and methylation sites (92%), conversely, a single cell study previously described a more complex relation depending on the location of CpG islands [43]. In addition, the analyses showed a general trend for metabolite candidates being positively correlated with gene expression, in agreement with previous literature [44].

CpGs in common between the methylation- and metabolite-driven integrations of omics belong to genes (*DHCR24* and *SC4MOL*) involved in cholesterol biosynthesis. Furthermore, C21-steroid hormone biosynthesis and metabolism was identified as an important metabolic pathway both through our methylation-driven analysis and in direct association with birthweight by Robinson et al. [8]. We therefore performed a verification study in the ENVIRONAGE cohort and showed that plasmalogen PC(36:4)/PC(O-36:5) was positively associated with HDL cholesterol levels that in turn were positively associated with higher birthweight. Phosphatidylcholine metabolism, and in particular plasmalogens, may regulate several important cholesterol biosynthesis processes which improve cholesterol sensing and facilitate interorganelle cholesterol trafficking [45]. During fetal development cholesterol and phospholipids are needed to build membranes, to develop the central nervous system including the brain and they are precursors of bile acids and steroid hormones [46]. While in adults LDLs

Table 4
Results from the cholesterol analyses.

	Total cholesterol		HDL cholesterol		LDL cholesterol	
	Change (mg/dl)	P-value	Change (mg/dl)	P-value	Change (mg/dl)	P-value
Metabolites[a]	*n* = 182		*n* = 182		*n* = 182	
PC(34:2)	17·08	**7·82e−04**	2·66	0·16	−1·07	0·48
Plasmalogen PC (36:4)/PC (O-36:5)	26·85	**3·05e−07**	5·28	**0·01**	−1·17	0·48
U61	17·70	**3·32e−04**	1·14	0·54	1·72	0·24
CpGs[a]	*n* = 178–179		n = 178–179		*n* = 178–179	
cg05119988	6·02	**7·45e−03**	0·55	0·51	0·17	0·80
cg17901584	6·87	**1·56e−03**	0·86	0·29	0·58	0·38
Proteins[a,b]	n = 176		*n* = 176		*n* = 176	
CCL22	−5·39	0·25	−2·88	0·09	0·90	0·51
Periostin	8·19	0·09	−2·70	0·13	0·89	0·52

P-values <0.05 are marked in bold. CCL22 = macrophage-derived chemokine; change = change in cholesterol levels (in mg/dl) for one interquartile range increment of metabolites, CpGs and proteins; HDL = high-density lipoprotein; LDL = low-density lipoprotein; n = numbers of observation; PC = phosphatidylcholine; U61 = unassigned metabolite 61.

[a] All the analyses are adjusted for gestational age, newborn sex, maternal age, maternal height, maternal BMI, smoking during pregnancy, parity, maternal education, and total cholesterol levels (for HDL and LDL analyses).

[b] Analyses of proteins were additionally adjusted on plate and analyses of CpGs were additionally adjusted for chip, position and cell types composition.

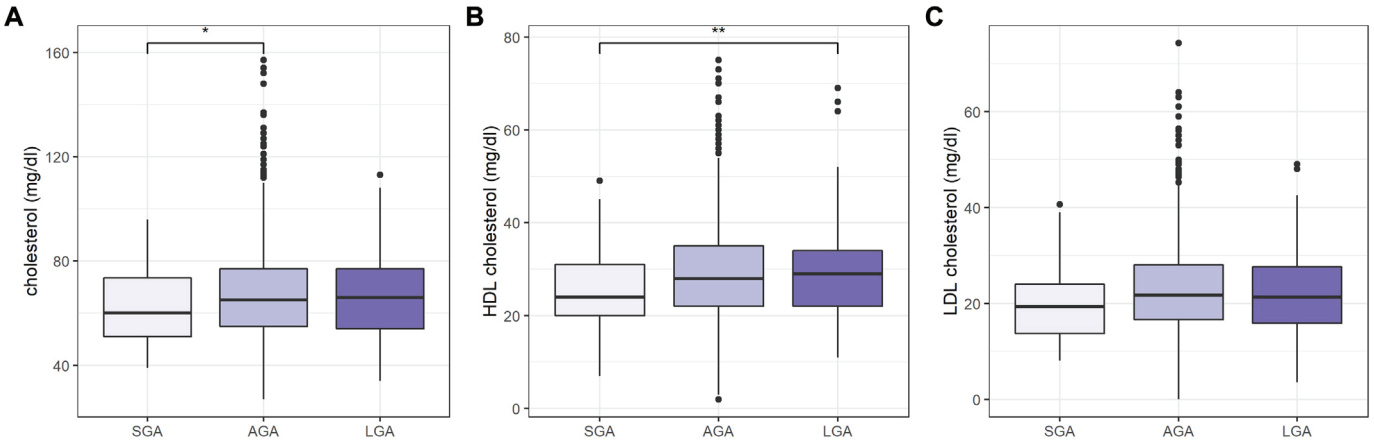


Fig. 6. Cholesterol levels in ENVIRONAGE cohort. A Total, B HDL and C LDL cholesterol levels in cord blood (on the y-axis) in SGA, AGA and LGA newborns (on the x-axis) of the ENVIRONAGE cohort ($n = 1097$) are graphically represented by boxplots. AGA = adequate for gestational age; LGA = large for gestational age; SGA = small for gestational age. *p-value between SGA and AGA and ** between SGA and LGA from linear multivariate analysis adjusted for gestational age, newborn sex, maternal age, maternal height, maternal BMI, maternal smoking during pregnancy, parity, maternal education, and cord blood total cholesterol levels (for HDL and LDL analyses) < 0·05.

are the major plasma lipoproteins, at birth cord blood is richer in HDLs because HDLs are produced in blood circulation and are not dependent upon fetal liver production, conversely to LDLs [47]. In our study the positive association between birthweight and HDL cholesterol was further confirmed by the finding of SGA having decreased HDL cholesterol levels compared to LGA newborns. Further, we also found decreased total cholesterol levels in SGA compared to both AGA and LGA newborns. Inconsistent (mostly null) associations of cord total and HDL cholesterol levels with birthweight and between SGA and AGA, that have been previously reported in literature [48–55], may be due to limited sample size or lack of adequate control regarding confounding. Conversely, positive associations between birthweight and HDL cholesterol, in line with our results, have been described by a randomized control trial in 343 obese pregnant women and an observational study in 1522 newborns upon adjustment for main confounders [16,56]. Further, this last study found lower total and HDL cholesterol levels in 105 SGA compared to 1320 AGA newborns. Beyond the traditional association of increased cardiovascular risk with concentrations of total and HDL cholesterol, HDL may have beneficial or detrimental effects on systemic inflammation, obesity and diabetes and aging depending on composition of HDL particles [57].

In the ENVIRONAGE cohort we explored if being SGA was a significant predictor of the omic markers identified in our study. Methylation levels of *cg05119988* and *cg17901584* located on *SC4MOL* and *DHCR24* genes were higher in SGA compared to LGA, and methylation of *cg17901584* was lower in LGA compared to AGA. Cord blood levels of the two CpGs have been previously associated with birthweight [9], but never before to SGA. In our study the methylation of these CpGs was further positively associated with total level of cord blood cholesterol and positively correlated with plasmalogen PC(36:4)/PC(O-36:5) and PC(34:2). Interestingly, previous research has associated methylation of *cg17901584* with waist circumference, PC(36:5) C and with HDL cholesterol in adults [58–60].

Our analysis had a number of weaknesses. Cord blood includes a mixture of cell-types that may demonstrate similar phenotypes but with distinct methylation and gene expression patterns [61,62]. The protein-set was limited to inflammatory proteins ($n = 16$) and the metabolome analysis was limited to a single analytical platform with many metabolites lacking annotation, which is common in metabolomics analyses [63]. Also, the curation of human pathway databases is incomplete and possibly biased towards specific diseases. Although we were not able to analyze genomic data in the same samples, we did not find in silico evidence of genetic variants influencing the DNA methylation sites. While we hypothesized that birthweight influences biomarkers at different omic levels, the cross-sectional study design does not allow assessment of causality and therefore we cannot exclude the possibility that the biomarkers themselves are responsible for birthweight modifications. In this regard, a recent multi-omic study in adults found through mendelian randomization that methylation of one of the two CpGs common to our methylation- and metabolite-driven integrations of omics (*cg17901584*) seems to be a consequence rather than a cause of obesity [64]. In sensitivity analyses we found differences by sex but we were not able to detect a clear pattern.

The major strengths of our study are the combination of cord blood samples from four different European birth-cohorts and an integrative analysis of different omic levels accompanied by computational and technical challenges. We acknowledge that our results may be affected by differences in birthweight across individual cohorts (Table 1), however the top signals in the two approaches were unaffected by factors differing between cohorts, indicating that study heterogeneity is not the main driver of our findings. Different methods have been described to integrate data obtained from different omic levels [31,65]. Our approach combines multi-omics correlation with pathway and network correlation analysis and aimed at identifying the intermediate biological mechanisms that link birthweight-related omic signatures from the same individuals.

The translational potential of our results lays in the development of an extensive catalog of birthweight associated signals. In observational studies, birthweight has repeatedly been associated with a variety of later-life diseases [4–7]. In the context of the DOHaD, the signals we found may reflect biomolecular changes exhibiting possible health effects later in life. For example, among the key signatures we identified, lower levels of cord blood PCs had been recently associated with higher risk of pulmonary hypertension in infants [66] and cord blood CCL22 has been associated with IgE sensitization in two year-old children [67]. Further, the finding of low levels of total and HDL cholesterol in SGA compared to AGA and LGA respectively, and that methylation of CpGs identified in our multi-omic study differed in SGA compared to LGA suggest a possible route for tailored intervention in SGA newborns that have higher risk of morbidity and mortality both in the perinatal period and in later life [68]. Although multiple molecular layers are linked via complex mechanisms, multi-omics integration, such as in our study, can further clarify relations between the different omics and enable us to study early life dynamics of molecular signals. Before we can translate this knowledge into general applications, the causality of the identified associations should be studied by longitudinal and in vivo experimental studies.

In conclusion, we further substantiated previously identified biomarkers in cord blood linked to birthweight and identified new omic features. Our data suggested that cholesterol and related metabolic pathways are related to birthweight. Our results provide evidence that

integration of different omic layers is a useful tool to generate hypotheses on mechanistic pathways. Further studies are required to discover the role of these biomarkers in later life diseases.

Funding information

This work is supported by the Bijzonder Onderzoeksfonds (BOF) Hasselt University through a PhD fellowship [to RA], the "EXPOsOMICS" grant [grant number 308610-FP7 European Commission to PV], and the "STOP" grant [grant number 774548-European Commission H2020 to PV]. The ENVIRONAGE birth-cohort is supported by the EU Program "Ideas" (ERC-2012-StG-310898) and the FWO (G082317N). Piccolipiù cohort has been funded by the CCM grant 2010 and the Italian Ministry of Health.

Disclaimer

Where authors are identified as personnel of the International Agency for Research on Cancer/World Health Organization, the authors alone are responsible for the views expressed in this article and they do not necessarily represent the decisions, policy or views of the International Agency for Research on Cancer/World Health Organization.

Availability of data and materials

EXPOsOMICS data analyzed during the current study are available via NCBI Gene Expression Omnibus (GEO) repository with the Accession No GSE151042 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE151042, for the methylome) and GSE151373 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE151373, for the transcriptome), and via the MetaboLights repository with the Accession No MTBLS1684 (https://www.ebi.ac.uk/metabolights/MTBLS1684, for the metabolome). Code relevant to the analyses is available upon request to the corresponding author.

CRediT authorship contribution statement

Rossella Alfano: Conceptualization, Formal analysis, Writing - original draft, Writing - review & editing. **Marc Chadeau-Hyam:** Conceptualization, Writing - review & editing. **Akram Ghantous:** Conceptualization, Data curation, Writing - review & editing. **Pekka Keski-Rahkonen:** Data curation, Writing - review & editing. **Leda Chatzi:** Resources, Writing - review & editing. **Almudena Espin Perez:** Writing - review & editing. **Zdenko Herceg:** Writing - review & editing. **Manolis Kogevinas:** Writing - review & editing. **Theo M. de Kok:** Data curation, Writing - review & editing. **Tim S. Nawrot:** Resources, Writing - review & editing. **Alexei Novoloaca:** Data curation, Writing - review & editing.**Chirag J. Patel:**Writing - review & editing. **Costanza Pizzi:** Resources, Writing - review & editing. **Nivonirina Robinot:** Data curation, Writing - review & editing. **Franca Rusconi:** Writing - review & editing. **Augustin Scalbert:** Writing - review & editing. **Jordi Sunyer:** Writing - review & editing. **Roel Vermeulen:** Data curation, Writing - review & editing. **Martine Vrijheid:** Resources, Writing - review & editing. **Paolo Vineis:** Conceptualization, Supervision, Writing - review & editing. **Oliver Robinson:** Conceptualization, Formal analysis, Writing - review & editing. **Michelle Plusquin:** Conceptualization, Writing - original draft, Writing - review & editing.

Declaration of competing interest

None.

Acknowledgments

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.metabol.2020.154292.

References

[1] Gluckman PD, Hanson MA, Gluckman P, Hanson M. The developmental origins of health and disease: an overview. In: Hanson M, Gluckman P, editors. Developmental origins of health and disease. Cambridge: Cambridge University Press; 2006. p. 1–5.

[2] Yokoyama Y, Jelenkovic A, Hur YM, Sund R, Fagnani C, Stazi MA, et al. Genetic and environmental factors affecting birth size variation: a pooled individual-based analysis of secular trends and global geographical differences using 26 twin cohorts. Int J Epidemiol. 2018;47:1195–206.

[3] Bianco-Miotto T, Craig JM, Gasser YP, van Dijk SJ, Ozanne SE. Epigenetics and DOHaD: from basics to birth and beyond. J Dev Orig Health Dis. 2017;8:513–9.

[4] Lawlor DA, Ronalds G, Clark H, Smith GD, Leon DA. Birth weight is inversely associated with incident coronary heart disease and stroke among individuals born in the 1950s: findings from the Aberdeen Children of the 1950s prospective cohort study. Circulation. 2005;112:1414–8.

[5] O'Donnell KJ, Meaney MJ. Fetal origins of mental health: the developmental origins of health and disease hypothesis. Am J Psychiatry. 2017;174:319–28.

[6] McCormack VA, dos Santos Silva I, Koupil I, Leon DA, Lithell HO. Birth characteristics and adult cancer incidence: Swedish cohort of over 11,000 men and women. Int J Cancer. 2005;115:611–7.

[7] Risnes KR, Vatten LJ, Baker JL, Jameson K, Sovio U, Kajantie E, et al. Birthweight and mortality in adulthood: a systematic review and meta-analysis. Int J Epidemiol. 2011;40:647–61.

[8] Robinson O, Keski-Rahkonen P, Chatzi L, Kogevinas M, Nawrot T, Pizzi C, et al. Cord blood metabolic signatures of birth weight: a population-based study. J Proteome Res. 2018;17:1235–47.

[9] Kupers LK, Monnereau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 2019;10:1893.

[10] Adkins RM, Tylavsky FA, Krushkal J. Newborn umbilical cord blood DNA methylation and gene expression levels exhibit limited association with birth weight. Chem Biodivers. 2012;9:888–99.

[11] Makikallio K, Kaukola T, Tuimala J, FK S, Hallman M, Ojaniemi M. Umbilical artery chemokine CCL16 is associated with preterm preeclampsia and fetal growth restriction. Cytokine. 2012;60:377–84.

[12] Engel SM, Joubert BR, Wu MC, Olshan AF, Haberg SE, Ueland PM, et al. Neonatal genome-wide methylation patterns in relation to birth weight in the Norwegian Mother and Child Cohort. Am J Epidemiol. 2014;179:834–42.

[13] Gillberg L, Perfilyev A, Brons C, Thomasen M, Grunnet LG, Volkov P, et al. Adipose tissue transcriptomics and epigenomics in low birthweight men and controls: role of high-fat overfeeding. Diabetologia. 2016;59:799–812.

[14] Perng W, Rifas-Shiman SL, McCulloch S, Chatzi L, Mantzoros C, Hivert MF, et al. Associations of cord blood metabolites with perinatal characteristics, newborn anthropometry, and cord blood hormones in project viva. Metabolism. 2017;76:11–22.

[15] Hellmuth C, Uhl O, Standl M, Demmelmair H, Heinrich J, Koletzko B, et al. Cord blood metabolome is highly associated with birth weight, but less predictive for later weight development. Obes Facts. 2017;10:85–100.

[16] Patel N, Hellmuth C, Uhl O, Godfrey K, Briley A, Welsh P, et al. Cord metabolic profiles in obese pregnant women: insights into offspring growth and body composition. J Clin Endocrinol Metab. 2018;103:346–55.

[17] Kadakia R, Talbot O, Kuang A, Bain JR, Muehlbauer MJ, Stevens RD, et al. Cord blood metabolomics: association with newborn anthropometrics and C-peptide across ancestries. J Clin Endocrinol Metab. 2019;104:4459–72.

[18] Lu YP, Reichetzeder C, Prehn C, Yin LH, Yun C, Zeng S, et al. Cord blood lysophosphatidylcholine 16: 1 is positively associated with birth weight. Cell Physiol Biochem. 2018;45:614–24.

[19] Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. Hum Mol Genet. 2015;24:3752–63.

[20] Agha G, Hajj H, Rifas-Shiman SL, Just AC, Hivert MF, Burris HH, et al. Birth weight-for-gestational age is associated with DNA methylation at birth and in childhood. Clin Epigenetics. 2016;8:118.

[21] Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, et al. The exposome in practice: design of the EXPOsOMICS project. Int J Hyg Environ Health. 2017;220:142–51.

[22] Janssen BG, Madhloum N, Gyselaers W, Bijnens E, Clemente DB, Cox B, et al. Cohort profile: the ENVIRonmental influence ON early AGEing (ENVIRONAGE): a birth cohort study. Int J Epidemiol. 2017;46:1386 [7m].

[23] Guxens M, Ballester F, Espada M, Fernandez MF, Grimalt JO, Ibarluzea J, et al. Cohort profile: the INMA–INfancia y Medio Ambiente–(environment and childhood) project. Int J Epidemiol. 2012;41:930–40.

[24] Farchi S, Forastiere F, Vecchi Brumatti L, Alviti S, Arnofi A, Bernardini T, et al. Piccolipiu, a multicenter birth cohort in Italy: protocol of the study. BMC Pediatr. 2014;14:36.

[25] Chatzi L, Leventakou V, Vafeiadi M, Koutra K, Roumeliotaki T, Chalkiadaki G, et al. Cohort profile: the mother-child cohort in Crete, Greece (Rhea Study). Int J Epidemiol. 2017;46:1392 [3k].

[26] Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis: chemical analysis working group (CAWG) metabolomics standards initiative (MSI). Metabolomics. 2007;3:211–21.
[27] Plusquin M, Chadeau-Hyam M, Ghantous A, Alfano R, Bustamante M, Chatzi L, et al. DNA methylome marks of exposure to particulate matter at three time points in early life. Environ Sci Technol. 2018;52:5427–37.
[28] Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, LM S, et al. DNA methylation of cord blood cell types: applications for mixed cell birth studies. Epigenetics. 2016; 11:354–62.
[29] Lubin JH, Colt JS, Camann D, Davis S, Cerhan JR, Severson RK, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ Health Perspect. 2004;112:1691–6.
[30] Patel CJ, Manrai AK. Development of exposome correlation globes to map out environment-wide associations. Pac Symp Biocomput. 2015;20:231–42.
[31] Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: viewpoints from the metabolomics research community. Metabolites. 2019;9.
[32] Bukowski R, Smith GC, Malone FD, Ball RH, Nyberg DA, Comstock CH, et al. Human sexual size dimorphism in early pregnancy. Am J Epidemiol. 2007;165:1216–8.
[33] Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting network activity from high throughput metabolomics. PLoS Comput Biol. 2013;9.
[34] Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018; 46 [W486-W94].
[35] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 2019;47 [D1005-D12].
[36] Rapp M, Wintergerst MWM, Kunz WG, Vetter VK, Knott MML, Lisowski D, et al. CCL22 controls immunity by promoting regulatory T cell communication with dendritic cells in lymph nodes. J Exp Med. 2019;216:1170–81.
[37] Anandi VS, Shaila B. Evaluation of factors associated with elevated newborn 17-hydroxyprogesterone levels. J Pediatr Endocrinol Metab. 2017;30:677–81.
[38] Lee JH, Ulrich B, Cho J, Park J, Kim CH. Progesterone promotes differentiation of human cord blood fetal T cells into T regulatory cells but suppresses their differentiation into Th17 cells. J Immunol. 2011;187:1778–87.
[39] Trinchieri G. Interleukin-12 and the regulation of innate resistance and adaptive immunity. Nat Rev Immunol. 2003;3:133–46.
[40] Safford MG, Levenstein M, Tsifrina E, Amin S, Hawkins AL, Griffin CA, et al. JAK3: expression and mapping to chromosome 19p12-13.1. Exp Hematol. 1997;25:374–86.
[41] Seif F, Khoshmirsafa M, Aazami H, Mohsenzadegan M, Sedighi G, Bahar M. The role of JAK-STAT signaling pathway and its regulators in the fate of T helper cells. Cell Commun Signal. 2017;15:23.
[42] Hellmuth C, Lindsay KL, Uhl O, Buss C, Wadhwa PD, Koletzko B, et al. Maternal metabolomic profile and fetal programming of offspring adiposity: identification of potentially protective lipid metabolites. Mol Nutr Food Res. 2019;63:e1700889.
[43] Hu Y, Huang K, An Q, Du G, Hu G, Xue J, et al. Simultaneous profiling of transcriptome and DNA methylome from a single cell. Genome Biol. 2016;17:88.
[44] Cuperlovic-Culf M, Ferguson D, Culf A, Morin Jr P, Touaibia M. 1H NMR metabolomics analysis of glioblastoma subtypes: correlation between metabolomics and gene expression characteristics. J Biol Chem. 2012;287:20164–75.
[45] Honsho M, Abe Y, Fujiki Y. Dysregulation of plasmalogen homeostasis impairs cholesterol biosynthesis. J Biol Chem. 2015;290:28822–33.
[46] Woollett LA. Review: transport of maternal cholesterol to the fetal circulation. Placenta. 2011;32(Suppl. 2):S218–21.
[47] Nagasaka H, Chiba H, Kikuta H, Akita H, Takahashi Y, Yanai H, et al. Unique character and metabolism of high density lipoprotein (HDL) in fetus. Atherosclerosis. 2002; 161:215–23.
[48] Brittos T, de Souza WB, Anschau F, Pellanda L. Lipids and leukocytes in newborn umbilical vein blood, birth weight and maternal body mass index. J Dev Orig Health Dis. 2016;7:672–7.
[49] Aletayeb SMH, Dehdashtian M, Aminzadeh M, Moghaddam A-RE, Mortazavi M, Malamiri RA, et al. Correlation between umbilical cord blood lipid profile and neonatal birth weight. Pediatr Pol. 2013;88:521–5.
[50] Ghiasi A, Ziaei S, Faghihzadeh S. The relationship between levels of lipids and lipoprotein B-100 in maternal serum and umbilical cord serum and assessing their effects on newborn infants anthropometric indices. Journal of Midwifery and Reproductive Health. 2014;2:227–32.
[51] Kenchappa Y, Behera N. Assay of neonatal cord blood lipid levels and its correlation with neonatal gestational age, gender and birth weight: a single center experience. International Journal of Contemporary Pediatrics. 2016;3:718–24.
[52] Nayak CD, Agarwal V, Nayak DM. Correlation of cord blood lipid heterogeneity in neonates with their anthropometry at birth. Indian J Clin Biochem. 2013;28:152–7.
[53] Katragadda T, Mahabala RS, Shetty S, Baliga S. Comparison of cord blood lipid profile in preterm small for gestational age and appropriate for gestational age newborns. J Clin Diagn Res. 2017;11:SC05–SC7.
[54] Hou RL, Jin WY, Chen XY, Jin Y, Wang XM, Shao J, et al. Cord blood C-peptide, insulin, HbA1c, and lipids levels in small- and large-for-gestational-age newborns. Med Sci Monit. 2014;20:2097–105.
[55] Lobo LL, Kumar HU, Mishra T, Sundari T, Singh A, Kumar CV, et al. Small-for-gestational-age versus appropriate-for-gestational-age: comparison of cord blood lipid profile & insulin levels in term newborns (SAGA-ACT study). Indian J Med Res. 2016;144:194–9.
[56] Wang J, Shen S, Price MJ, Lu J, Sumilo D, Kuang Y, et al. Glucose, insulin, and lipids in cord blood of neonates and their association with birthweight: differential metabolic risk of large for gestational age and small for gestational age babies. J Pediatr. 2020; 220 64-72.e2.
[57] Hafiane A, Favari E, Daskalopoulou SS, Vuilleumier N, Frias MA. High-density lipoprotein cholesterol efflux capacity and cardiovascular risk in autoimmune and non-autoimmune diseases. Metabolism. 2020;104:154141.
[58] Petersen AK, Zeilinger S, Kastenmuller G, Romisch-Margl W, Brugger M, Peters A, et al. Epigenetics meets metabolomics: an epigenome-wide association study with blood serum metabolic traits. Hum Mol Genet. 2014;23:534–45.
[59] Demerath EW, Guan W, Grove ML, Aslibekyan S, Mendelson M, Zhou YH, et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. Hum Mol Genet. 2015;24:4464–79.
[60] Braun KVE, Dhana K, de Vries PS, Voortman T, van Meurs JBJ, Uitterlinden AG, et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam study. Clin Epigenetics. 2017;9:15.
[61] Adalsteinsson BT, Gudnason H, Aspelund T, Harris TB, Launer LJ, Eiriksdottir G, et al. Heterogeneity in white blood cells has potential to confound DNA methylation measurements. PLoS One. 2012;7:e46705.
[62] Xu Q, Ni S, Wu F, Liu F, Ye X, Mougin B, et al. Investigation of variation in gene expression profiling of human blood by extended principle component analysis. PLoS One. 2011;6:e26905.
[63] Sun YV, Hu YJ. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet. 2016;93:147–90.
[64] Liu J, Carnero-Montoro E, van Dongen J, Lent S, Nedeljkovic I, Ligthart S, et al. An integrative cross-omics analysis of DNA methylation sites of glucose and insulin homeostasis. Nat Commun. 2019;10:2581.
[65] Perakakis N, Yazdani A, Karniadakis GE, Mantzoros C. Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. Metabolism. 2018;87:A1–9.
[66] La Frano MR, Fahrmann JF, Grapov D, Pedersen TL, Newman JW, Fiehn O, et al. Umbilical cord blood metabolomics reveal distinct signatures of dyslipidemia prior to bronchopulmonary dysplasia and pulmonary hypertension. Am J Physiol Lung Cell Mol Physiol. 2018;315 [L870-L81].
[67] Yeh KW, Chiu CY, Su KW, Tsai MH, Hua MC, Liao SL, et al. High cord blood CCL22/CXCL10 chemokine ratios precede allergic sensitization in early childhood. Oncotarget. 2017;8:7384–90.
[68] Saenger P, Czernichow P, Hughes I, Reiter EO. Small for gestational age: short stature and beyond. Endocr Rev. 2007;28:219–51.

VII. Discussion

A. Methodology

This thesis investigated several different types of omic data with emphasis on population-based DNA methylome-wide studies. Generation, analysis and interpretation of this type of data are not straightforward (163,164). For this reason, we have invested in progressively optimizing existing bioinformatic and biostatistic pipelines for methylation data. This includes at the same time small but essential improvements as well as more extensive collaborative investigations. The first paper of the methods (144) focused on identifying and removing unwanted sources of variation such as batch effects which can be sometimes overlooked but may cause spurious results especially if they correlate with biological variables being studied (165). The paper compared the performance of several batch correction techniques on 902 samples from European Prospective Investigation into Cancer and Nutrition (EPIC) study (166). We concluded that the SVA approach outperforms the other methods tested for batch correction especially that, unlike the other methods, SVA estimates latent variables, thus, does not require the sources of variability to be known. The recent advances in high-throughput technologies create an important necessity for such or newer methods, and it is critical to continuously monitor and benchmark these types of approaches in light of the rapid evolvement of the omics field.

The second methodology paper benchmarked six popular integrative clustering methods using data simulation and publicly available omics data from breast cancer tissues. The results from the simulations and application revealed that matrix factorization methods were generally better to identify shared variation across several omic datasets. Sample clustering constituted the main evaluation criterion; however, more complete comparison could be done by also assessing variable clustering performance. Most of the investigated methods offer the possibility to do feature selection. It would be worth investigating the impact of penalization techniques on the method performances. All our tested simulations also demonstrated an advantage of integrative over non-integrative methods in the identification of common structure, supporting their use in the identification of complex structures across omic layers and in complex diseases such as cancer.

The third paper represents a pan-cancer investigation using publicly available data from The Cancer Genome Atlas (TCGA). These consisted of datasets with genetic alterations and gene expression of 426 epigenetic regulator genes (ERGs) in 33 cancer types. The high frequency of genetic alterations in ERGs in common human cancers (87,167) constitutes a “genetic smoking gun” that epigenetic mechanisms lie at the very heart of cancer biology. This study contributes to a greater understanding of the deregulation of ERGs and their functional impact in cancer and should prove instrumental in the clinical application of ERGs. We developed statistical and bioinformatics tools to predict cancer driver potential of gene function and to disentangle driver from passenger genes from a plethora of differentially regulated molecular players in cancer, most of which are often a resultant of the cancer rather

than a (driver) cause to the cancer. We aim in follow-up studies to this work to test these tools on the epigenetic markers identified herein in order to characterize their potential diver roles in childhood carcinogenesis.

B. Early-life factors, DNA methylation and childhood cancer

The acquired methodology expertise was used for downstream analysis of DNA methylation and specific early-life factors, being BW, gestational age and child sex, which represents tightly linked intrinsic factors. Each of the three was investigated in some of the largest studies to date, and the resultant findings were analysed in relation to childhood cancer risk through a proposed three-way modelling approach (in preparation). To complement these findings, a hypothesis-free investigation was also performed, but statistical power and robustness of findings yielded from such an agnostic methylome-wide approach are limited, requiring further replication in larger sample sizes and additional cohorts as well as adjustment for potential confounders.

1. Birthweight

BW was investigated using a meta-analysis of epigenome-wide association studies of 8,825 neonates from 24 birth cohorts in PACE consortium. We showed that DNA methylation in neonatal blood was associated with BW at 914 CpG sites (Bonferroni p.value < 0.05), with a difference in BW ranging from -183 to 178 grams per 10% increase in methylation. Pathway enrichment analysis of the significant CpGs involved in transcription regulation and skeletal and blood system development. The 914 neonatal blood CpGs were examined in follow-up studies in older children. A small fraction (1.3%) of birthweight-associated CpGs remain differentially methylated in childhood and adolescence. Previous studies reported similar rapid attenuation of differential methylation in relation to BW in the first years after birth (168). Large longitudinal studies would explore persistence at older ages in more detail and with more power. That said, persistence of differential methylation may not be imperative in predisposing to later adverse health effects as specific epigenetic events during a critical developmental period could initiate a program which later in life could be important, regardless of the continued presence or absence of that initiator (a “hit and run” effect). We used the adult whole blood reference for estimating the WBC proportions because cohort-specific analyses were completed before the widespread use of the cord blood reference became available (149). However, we found similar results after rerunning a sensitivity analysis on two cohorts with the new cord blood estimations.

We also undertook Mendelian randomization approach to explore potential causal associations with BW and later in life phenotypes using publicly available summary data. However, for the vast majority (86%) of the BW-associated CpGs, no genetic instrumental variables were identified; for 12%, only one instrumental variable was identified, and for the remaining 2%, none of the CpGs had more than 4 instrumental

variables. Having only one instrumental variable may often result in biases due to the horizontal pleiotropy (a single instrumental variable influencing multiple traits) and assumption of spurious causation. Although the MR results were not conclusive, we expect that future development of GWAS of EWAS databases could help tackle this problem.

Post hoc power calculation indicates that despite being the largest study to date on DNA methylation association with BW, small but relevant differences may not have been identified in the current analysis. However, this scientific publication may have important public health impact in children and represents an important step forward to achieving the thesis aims.

The newborn epigenetic biomarkers of BW were further investigated in a follow up study integrating four types of omic data: methylome, transcriptome, metabolome and a set of inflammatory proteins from four birth-cohorts from the EXPOsOMICS European project. Most of the identified signals across the different omic layers were related to cholesterol biosynthesis. This was further supported by significant associations between HDL cholesterol levels measured in cord blood and BW in the same cohorts. This study combines multi-omic integration skills acquired in the methodology paper and findings from the EWAS study on BW in an attempt to elucidate how cross-correlation between omic layers can further unravel biological pathways of birthweight.

2. Gestational age

The study in the second paper was based on 6,885 neonates from 20 cohorts worldwide and provided a comprehensive catalogue of DNA methylation markers of gestational age (range 27–42 weeks), a birth characteristic associated with later onset of diseases (169–174) including cancer (175). We identified 8899 CpGs in cord blood that were associated with pregnancy duration, at Bonferroni significance spanning all chromosomes after adjustment for confounders including cord blood estimated WBC. The largest association represents 2.5% methylation change per additional gestational week. The most significant positively differentially methylated CpG in cord blood was located in IGF2BP1, a gene known to be involved in adiposity and cardiometabolic disease risk (176), and to play an essential role in embryogenesis and carcinogenesis (177,178). Enrichment for biological pathways involved processes critical to development and related to autoimmune and inflammatory diseases. Given the large number of significant associations, we proceeded to a selection of CpGs via two different techniques: regions that had at least three adjacent significant CpGs and differentially methylated region (DMR) analysis. The latter identified more than 95% of the 1276 CpGs selected using at least three consecutive sites. These 1276 CpGs annotated to 325 genes and were used in the downstream analyses. We acknowledge that applying this filter may have led to potentially important single CpGs not being included in the functional analyses. In addition, genes with few CpGs represented on the 450K array are likely under-represented in the downstream analyses.

We performed look-up analyses in older ages: early childhood (4–5 years), school age (7–9 years) and adolescence (16–18 years) in a total of 2481 children. Out of the 1276 CpGs, we observed 40, 60 and 60 sites in the three age groups respectively to be associated with gestational age at the nominal significance level (p.value < 0.05) with the same direction of effect. One CpG located in TMEM176B gene and previously associated with gestational age at birth (179) was nominally significant in all age groups with same direction of effect. The protein encoded by this gene has been proposed as a potential biomarker for various cancers (180). The low number of significant sites which, in addition did not survive multiple test correction, may be partially explained by smaller sample sizes in children compared to newborn analyses and by the fact that many later exposures may obscure the association.

We also completed a longitudinal analysis using DNA methylation from two time points (birth and 4 years) in one cohort and three times points (birth, 7 and 17 years) in another cohort. Blood methylation levels at most identified CpGs changed significantly during early childhood with stabilization at school age. However, a subset (17%) of stable CpGs changed little from birth to adolescence.

In addition to the lookup and longitudinal analyses in blood over time, we investigated whether the CpGs detected in cord blood (which originates from the mesoderm embryonic layer) were differentially methylated in relation to gestational age in other fetal tissues, lung and brain (which originate from the two other embryonic germ layers, ectoderm and endoderm, respectively) (181). We found clear overlap of methylation markers, highlighting that the cord blood findings capture the epigenomic plasticity of pre-natal development across tissues.

To further investigate a potential functional impact of methylation changes on gene expression in cord blood (with focus on cis-effects), publicly available datasets were used leading to the identification of multiple strong correlations between methylation and gene expression. Forty-six percent of these DNA methylation-expression correlations were negative and fifty-four percent were positive. These expression findings may reflect relevance for health outcomes associated with gestation age.

This large-scale study expands our knowledge on newborn methylation differences in relation to pregnancy duration, some of them being observed later in childhood. The availability of samples at multiple ages and our ability to compare our findings with those in fetal tissue datasets represent major strengths of the study. This scientific article represents an essential element towards the achievement of the thesis objectives.

3. Child sex

It is known that many cancers exhibit a gender-bias and CC is not an exception. We have also shown that gender significantly influences the interaction between birth order and childhood leukemia risk (78) as well as the interaction between the methylome and childhood cancer (in preparation). Hence, we also completed an analysis aiming to decipher how autosomal methylation patterns dictate or are affected by sex (in addition to the well-characterized sex chromosome methylation

patterns). The meta-analysis of child sex and offspring DNA methylation association at birth was performed using 8314 cord blood samples from 16 cohorts within PACE consortium. Over 40,000 CpGs sites of the 450K methylation array were differentially methylated with small but statistically significant differences between boys and girls at birth. In the lookup analyses, most of the signals persisted in older ages with similar directions of effect. The significant CpGs were enriched in cancer pathways and neurological disorders in both newborns and children. Our findings proved an identical trend with a previous study (182) for hypomethylation (67%) in boys for both autosomal and X chromosome sites. These results suggest that DNA methylation may contribute to developmental differences between boys and girls that impact sex-dependent differences in health.

There are several strengths and limitations of this study. The association between child sex and DNA methylation was investigated using robust statistical models in the largest studies to date ensuring enough statistical power to assess small effect sizes. WBC composition was estimated using cord blood reference for newborns and adult whole blood reference for children and adjusted for in the different models. Having two distinct time-points allowed to investigate the stability in time of the differentially methylated CpGs found in newborns. Our study did not investigate if the methylation changes are impacting gene expression. Thus, follow-up studies are required to confirm if these methylation differences extend to functional changes in order to confirm the biological significance and contribution towards the fetal origins of disease hypothesis.

4. Bringing it all together

The three major studies on the identification of epigenetic markers of BW, gestational age and child sex incur some common limitations. Early-life factors and DNA methylation were measured at the same time point. Thus, causality and its direction are difficult to ascertain, despite the fact that we adjusted for potential confounders, including WBC composition. As the majority of the participants were of European ancestry, more studies involving a larger number of non-European samples are needed to ensure generalizability of results.

The relationship between the identified methylation markers of intrinsic factors and childhood cancer risk is currently being investigated through our three-way modelling approach. The preliminary results highlight one CpG in a non-coding gene that mediates the association between BW and CL and is not confounded by gestational age or child sex. The mediation was significant *via* two different and well-established statistical methods, but requires replication in additional cohorts and further experimental verification through functional assays.

We also agnostically investigated the association between DNA methylation and CC risk in a hypothesis-free approach. This has led to the identification of four differentially methylated genomic regions (each spanning 3 to 16 methylation sites). The CpGs obtained from the hypothesis-free approach were not enriched in BW markers, though this would require larger sample sizes to confirm and to adjust for potential interference from confounders, given that the agnostic investigation has

limited statistical power and was based on a small number of studies. The two approaches are complementary with a primary purpose to provide a more comprehensive coverage of significant molecular associations and are not expected to necessarily yield overlapping results.

C. Future perspectives

This project constitutes several steps, intricately linked in order to synthesize meaningful associations linking early-life factors, epigenetic mechanisms and CL risk. Ongoing efforts aim to additionally expand the cancer scope from leukemias to other less common forms, such as child brain cancers, especially that we have recently assembled larger sample sizes from various studies. Among other early-life factors, investigating extrinsic exposure factors would be a logical next step, and birth order is an interesting example to explore as it is one of the few and new prospectively established risk factors for CC. Birth order is often used as a proxy for early-life infection as it has been associated with an increased risk of common infections found in blood. Prospective evidence for pesticides indicates an increased risk for AML making it, along with birth order, interesting factors to be prioritized for future mechanistic studies.

Another approach would be to use existing epigenetic biomarkers of early-life factors with for which evidence is not yet very convincing, namely maternal smoking and air pollution. Both factors were studied in relation to DNA methylation by meta-analyzing the results of 13 and 9 cohorts respectively. Smoking had a profound impact on the epigenome with among the results AHRR (118), one of the best biomarkers of smoking, which unlike cotinine that have a very short half-life, can persist up to 30 years. Contrarily, air pollution showed a modest effect in association with DNA methylation for both particulate matter (PM) with diameter <10μm (PM10) and with diameter <2,5μm (PM2,5) (183). Furthermore, tobacco and air pollution are established carcinogens by IARC, and their relation to childhood cancer is planned to be investigated in the epidemiological arm of I4C. Three-way modelling approach may prove to be a useful tool to expand our work into more exposures, especially those for which we have already catalogued epigenetic markers in population-based studies (such as the case of tobacco smoking and air pollution). Finally, pediatric neoplasms could be scrutinized for “epidrivers” genes following the framework for adult cancers presented in the Methods section and using the cancer driver prediction tools we have developed.

VIII. Conclusion

We have extensively developed and optimized statistical and bioinformatics frameworks needed for the overall analyses of this thesis, including epigenomics batch corrections, meta-analysis of molecular epidemiology data and multi-omics data integration. The three inter-related intrinsic early-life factors were shown to have a profound association with the neonatal epigenome. These findings provide a comprehensive catalogue of differential methylation in relation to these early-life factors, which may prove additionally useful to the growing community of researchers studying DOHaD. Studying the dynamics of identified CpGs in childhood and adolescence demonstrated that some of the newborn signals persisted in older ages. Functional pathway enrichment demonstrated the involvement of these epigenetic markers in human diseases and biological processes critical to fetal development. Our findings highlight potential biological mechanisms that could underlie the associations between BW, its closely related intrinsic factors gestational age and child sex, and childhood leukemia. Replication of these findings and further in-depth functional analysis through experimental models may help ascertain some of the identified epigenetic biomarkers and characterize driver genes and causal pathways that are implicated in pediatric carcinogenesis. Such studies have the potential to enhance our knowledge of CC etiology and provide an evidence base for cancer prevention.

List of figures

Bibliography

1. Johnston WT, Erdmann F, Newton R, Steliarova-Foucher E, Schüz J, Roman E. Childhood cancer: Estimating regional and global incidence. Cancer Epidemiol. 7 janv 2020;101662.

2. ACCIS: The Automated Cancer Information System [Internet]. [cité 24 mars 2020]. Disponible sur: http://accis.iarc.fr/

3. Terracini B. Epidemiology of childhood cancer. Environ Health. 5 avr 2011;10(Suppl 1):S8.

4. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7-30.

5. Stewart BW, Wild C, International Agency for Research on Cancer, World Health Organization. World cancer report 2014. 2014.

6. Gupta S, Howard SC, Hunger SP, Antillon FG, Metzger ML, Israels T, et al. Treating Childhood Cancer in Low- and Middle-Income Countries. In: Gelband H, Jha P, Sankaranarayanan R, Horton S, éditeurs. Cancer: Disease Control Priorities, Third Edition (Volume 3) [Internet]. Washington (DC): The International Bank for Reconstruction and Development / The World Bank; 2015 [cité 28 févr 2020]. Disponible sur: http://www.ncbi.nlm.nih.gov/books/NBK343626/

7. Howard SC, Zaidi A, Cao X, Weil O, Bey P, Patte C, et al. The My Child Matters programme: effect of public-private partnerships on paediatric cancer care in low-income and middle-income countries. Lancet Oncol. 2018;19(5):e252-66.

8. Erdmann F, Ghantous A, Schüz J. Environmental Agents and Childhood Cancer. In: Nriagu J, éditeur. Encyclopedia of Environmental Health (Second Edition) [Internet]. Oxford: Elsevier; 2019 [cité 25 févr 2020]. p. 347-59. Disponible sur: http://www.sciencedirect.com/science/article/pii/B9780124095489117257

9. Ionizing radiation, health effects and protective measures [Internet]. [cité 28 févr 2020]. Disponible sur: https://www.who.int/news-room/fact-sheets/detail/ionizing-radiation-health-effects-and-protective-measures

10. Brenner DJ, Hall EJ. Computed tomography--an increasing source of radiation exposure. N Engl J Med. 29 nov 2007;357(22):2277-84.

11. El Ghissassi F, Baan R, Straif K, Grosse Y, Secretan B, Bouvard V, et al. A review of human carcinogens--part D: radiation. Lancet Oncol. août 2009;10(8):751-2.

12. Little MP, Wakeford R, Borrego D, French B, Zablotska LB, Adams MJ, et al. Leukaemia and myeloid malignancy among people exposed to low doses (<100

mSv) of ionising radiation during childhood: a pooled analysis of nine historical cohort studies. Lancet Haematol. août 2018;5(8):e346-58.

13. Little JB. Radiation carcinogenesis. Carcinogenesis. mars 2000;21(3):397-404.

14. Preston DL, Cullings H, Suyama A, Funamoto S, Nishi N, Soda M, et al. Solid Cancer Incidence in Atomic Bomb Survivors Exposed In Utero or as Young Children. JNCI J Natl Cancer Inst. 19 mars 2008;100(6):428-36.

15. Ozasa K, Shimizu Y, Suyama A, Kasagi F, Soda M, Grant EJ, et al. Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases. Radiat Res. mars 2012;177(3):229-43.

16. Yamashita S, Suzuki S, Suzuki S, Shimura H, Saenko V. Lessons from Fukushima: Latest Findings of Thyroid Cancer After the Fukushima Nuclear Power Plant Accident. Thyroid. 1 janv 2018;28(1):11-22.

17. Harbron RW, Feltbower RG, Glaser A, Lilley J, Pearce MS. Secondary malignant neoplasms following radiotherapy for primary cancer in children and young adults. Pediatr Hematol Oncol. avr 2014;31(3):259-67.

18. Armstrong GT, Liu Q, Yasui Y, Huang S, Ness KK, Leisenring W, et al. Long-term outcomes among adult survivors of childhood central nervous system malignancies in the Childhood Cancer Survivor Study. J Natl Cancer Inst. 1 juill 2009;101(13):946-58.

19. Inskip PD, Robison LL, Stovall M, Smith SA, Hammond S, Mertens AC, et al. Radiation Dose and Breast Cancer Risk in the Childhood Cancer Survivor Study. J Clin Oncol. 20 août 2009;27(24):3901-7.

20. Bhatia S, Sklar C. Second cancers in survivors of childhood cancer. Nat Rev Cancer. févr 2002;2(2):124-32.

21. Moskowitz CS, Chou JF, Wolden SL, Bernstein JL, Malhotra J, Novetsky Friedman D, et al. Breast cancer after chest radiation therapy for childhood cancer. J Clin Oncol Off J Am Soc Clin Oncol. 20 juill 2014;32(21):2217-23.

22. Giles D, Hewitt D, Stewart A, Webb J. Malignant disease in childhood and diagnostic irradiation in utero. Lancet Lond Engl. 1 sept 1956;271(6940):447.

23. Schulze-Rath R, Hammer GP, Blettner M. Are pre- or postnatal diagnostic X-rays a risk factor for childhood cancer? A systematic review. Radiat Environ Biophys. juill 2008;47(3):301-12.

24. McCollough CH, Schueler BA, Atwell TD, Braun NN, Regner DM, Brown DL, et al. Radiation exposure and pregnancy: when should we be concerned? Radiogr Rev Publ Radiol Soc N Am Inc. août 2007;27(4):909-17; discussion 917-918.

25. Bartley K, Metayer C, Selvin S, Ducore J, Buffler P. Diagnostic X-rays and risk of childhood leukaemia. Int J Epidemiol. déc 2010;39(6):1628-37.

26. Shu XO, Potter JD, Linet MS, Severson RK, Han D, Kersey JH, et al. Diagnostic X-rays and ultrasound exposure and risk of childhood acute lymphoblastic leukemia by immunophenotype. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. févr 2002;11(2):177-85.

27. Shu XO, Jin F, Linet MS, Zheng W, Clemens J, Mills J, et al. Diagnostic X-ray and ultrasound exposure and risk of childhood cancer. Br J Cancer. sept 1994;70(3):531-6.

28. Infante-Rivard C. Diagnostic x rays, DNA repair genes and childhood acute lymphoblastic leukemia. Health Phys. juill 2003;85(1):60-4.

29. Infante-Rivard C, Mathonnet G, Sinnett D. Risk of childhood leukemia associated with diagnostic irradiation and polymorphisms in DNA repair genes. Environ Health Perspect. juin 2000;108(6):495-8.

30. Pearce MS, Salotti JA, Little MP, McHugh K, Lee C, Kim KP, et al. Radiation exposure from CT scans in childhood and subsequent risk of leukaemia and brain tumours: a retrospective cohort study. Lancet. 4 août 2012;380(9840):499-505.

31. Mathews JD, Forsythe AV, Brady Z, Butler MW, Goergen SK, Byrnes GB, et al. Cancer risk in 680,000 people exposed to computed tomography scans in childhood or adolescence: data linkage study of 11 million Australians. BMJ. 21 mai 2013;346:f2360.

32. Little JB. Radiation carcinogenesis. Carcinogenesis. 1 mars 2000;21(3):397-404.

33. Morgan WF, Sowa MB. Non-targeted effects induced by ionizing radiation: mechanisms and potential impact on radiation induced health effects. Cancer Lett. 1 janv 2015;356(1):17-21.

34. Aypar U, Morgan WF, Baulch JE. Radiation-induced genomic instability: Are epigenetic mechanisms the missing link? Int J Radiat Biol. 1 févr 2011;87(2):179-91.

35. Pogribny I, Koturbash I, Tryndyak V, Hudson D, Stevenson SML, Sedelnikova O, et al. Fractionated low-dose radiation exposure leads to accumulation of DNA damage and profound alterations in DNA and histone methylation in the murine thymus. Mol Cancer Res MCR. oct 2005;3(10):553-61.

36. Pogribny I, Raiche J, Slovack M, Kovalchuk O. Dose-dependence, sex- and tissue-specificity, and persistence of radiation-induced genomic DNA methylation changes. Biochem Biophys Res Commun. 6 août 2004;320(4):1253-61.

37. Kalinich JF, Catravas GN, Snyder SL. The effect of gamma radiation on DNA methylation. Radiat Res. févr 1989;117(2):185-97.

38. Belinsky SA, Klinge DM, Liechty KC, March TH, Kang T, Gilliland FD, et al. Plutonium targets the p16 gene for inactivation by promoter hypermethylation in human lung adenocarcinoma. Carcinogenesis. juin 2004;25(6):1063-7.

39. WHO | Pesticides [Internet]. WHO. [cité 14 févr 2020]. Disponible sur: http://www.who.int/topics/pesticides/en/

40. IARC. Occupational Exposures in Insecticide Application, and Some Pesticides [Internet]. [cité 14 févr 2020]. Disponible sur: https://publications.iarc.fr/Book-And-Report-Series/Iarc-Monographs-On-The-Identification-Of-Carcinogenic-Hazards-To-Humans/Occupational-Exposures-In-Insecticide-Application-And-Some-Pesticides-1991

41. Zahm SH, Ward MH. Pesticides and childhood cancer. Environ Health Perspect. juin 1998;106(Suppl 3):893-908.

42. Simcox NJ, Fenske RA, Wolz SA, Lee IC, Kalman DA. Pesticides in household dust and soil: exposure pathways for children of agricultural families. Environ Health Perspect. déc 1995;103(12):1126-34.

43. Ostrea EM, Bielawski DM, Posecion NC, Corrion M, Villanueva-Uy E, Bernardo RC, et al. Combined analysis of prenatal (maternal hair and blood) and neonatal (infant hair, cord blood and meconium) matrices to detect fetal exposure to environmental pesticides. Environ Res. janv 2009;109(1):116-22.

44. Kaatsch P. Epidemiology of childhood cancer. Cancer Treat Rev. 1 juin 2010;36(4):277-85.

45. Van Maele-Fabry G, Lantin A-C, Hoet P, Lison D. Residential exposure to pesticides and childhood leukaemia: A systematic review and meta-analysis. Environ Int. 1 janv 2011;37(1):280-91.

46. Metayer C, Milne E, Clavel J, Infante-Rivard C, Petridou E, Taylor M, et al. The Childhood Leukemia International Consortium. Cancer Epidemiol. juin 2013;37(3):336-47.

47. Bailey HD, Infante-Rivard C, Metayer C, Clavel J, Lightfoot T, Kaatsch P, et al. Home pesticide exposures and risk of childhood leukemia: Findings from the Childhood Leukemia International Consortium. Int J Cancer J Int Cancer. 1 déc 2015;137(11):2644-63.

48. Bailey HD, Fritschi L, Infante-Rivard C, Glass DC, Miligi L, Dockerty JD, et al. Parental occupational pesticide exposure and the risk of childhood leukemia in the offspring: Findings from the childhood leukemia international consortium. Int J Cancer. 2014;135(9):2157-72.

49. Van Maele-Fabry G, Hoet P, Lison D. Parental occupational exposure to pesticides as risk factor for brain tumors in children and young adults: A systematic review and meta-analysis. Environ Int. 1 juin 2013;56:19-31.

50. Brown RC, Dwyer T, Kasten C, Krotoski D, Li Z, Linet MS, et al. Cohort Profile: The International Childhood Cancer Cohort Consortium (I4C). Int J Epidemiol. 1 août 2007;36(4):724-30.

51. Patel DM, Jones RR, Booth BJ, Olsson AC, Kromhout H, Straif K, et al. Parental occupational exposure to pesticides, animals and organic dust and risk of

childhood leukemia and central nervous system tumors: Findings from the International Childhood Cancer Cohort Consortium (I4C). Int J Cancer. 2020;146(4):943-52.

52. Herceg Z, Ghantous A, Wild CP, Sklias A, Casati L, Duthie SJ, et al. Roadmap for Investigating Epigenome Deregulation and Environmental Origins of Cancer. Int J Cancer. 1 mars 2018;142(5):874-82.

53. Zhang X, Wallace AD, Du P, Kibbe WA, Jafari N, Xie H, et al. DNA methylation alterations in response to pesticide exposure in vitro. Environ Mol Mutagen. août 2012;53(7):542-9.

54. Benedetti D, Lopes Alderete B, de Souza CT, Ferraz Dias J, Niekraszewicz L, Cappetta M, et al. DNA damage and epigenetic alteration in soybean farmers exposed to complex mixture of pesticides. Mutagenesis. 24 févr 2018;33(1):87-95.

55. Cedergreen N. Quantifying Synergy: A Systematic Review of Mixture Toxicity Studies within Environmental Toxicology. PLOS ONE. 2 mai 2014;9(5):e96580.

56. MacMahon B, Newill VA. Birth Characteristics of Children Dying of Malignant Neoplasms. JNCI J Natl Cancer Inst. 1 janv 1962;28(1):231-44.

57. Forsberg JG, Källén B. Pregnancy and delivery characteristics of women whose infants develop child cancer. A study based on registry information. APMIS Acta Pathol Microbiol Immunol Scand. janv 1990;98(1):37-42.

58. Harder T, Plagemann A, Harder A. Birth weight and risk of neuroblastoma: a meta-analysis. Int J Epidemiol. juin 2010;39(3):746-56.

59. Harder T, Plagemann A, Harder A. Birth weight and subsequent risk of childhood primary brain tumors: a meta-analysis. Am J Epidemiol. 15 août 2008;168(4):366-73.

60. Heck JE, Meyers TJ, Lombardi C, Park AS, Cockburn M, Reynolds P, et al. Case-control study of birth characteristics and the risk of hepatoblastoma. Cancer Epidemiol. août 2013;37(4):390-5.

61. Milne E, Greenop KR, Metayer C, Schüz J, Petridou E, Pombo-de-Oliveira MS, et al. Fetal Growth and Childhood Acute Lymphoblastic Leukemia: Findings from the Childhood Leukemia International Consortium (CLIC). Int J Cancer J Int Cancer. 15 déc 2013;133(12):2968-79.

62. Dahlhaus A, Prengel P, Spector L, Pieper D. Birth weight and subsequent risk of childhood primary brain tumors: An updated meta-analysis. Pediatr Blood Cancer. 2017;64(5):e26299.

63. Paltiel O, Tikellis G, Linet M, Golding J, Lemeshow S, Phillips G, et al. Birthweight and Childhood Cancer: Preliminary Findings from the International Childhood Cancer Cohort Consortium (I4C). Paediatr Perinat Epidemiol. juill 2015;29(4):335-45.

64. Lee J, Chia K-S, Cheung K-H, Chia S-E, Lee H-P. Birthweight and the risk of early childhood cancer among Chinese in Singapore. Int J Cancer. 2004;110(3):465-7.

65. Murray L, McCarron P, Bailie K, Middleton R, Davey Smith G, Dempsey S, et al. Association of early life factors and acute lymphoblastic leukaemia in childhood: historical cohort study. Br J Cancer. 1 févr 2002;86(3):356-61.

66. Westergaard T, Andersen PK, Pedersen JB, Olsen JH, Frisch M, Sørensen HT, et al. Birth characteristics, sibling patterns, and acute leukemia risk in childhood: a population-based cohort study. J Natl Cancer Inst. 2 juill 1997;89(13):939-47.

67. W.C.R.F.A.I.f.C. Diet, Nutrition, Physical activity and Cancer: a Global Perspective. Continuous Update Project Expert Third Report. 2018.

68. Felix JF, Joubert BR, Baccarelli AA, Sharp GC, Almqvist C, Annesi-Maesano I, et al. Cohort Profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. Int J Epidemiol. févr 2018;47(1):22-23u.

69. Küpers LK, Monnereau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun. 23 avr 2019;10(1):1-11.

70. Greaves M. The ‘delayed infection’ (aka ‘hygiene’) hypothesis for childhood leukaemia. In: Rook GAW, éditeur. The Hygiene Hypothesis and Darwinian Medicine [Internet]. Basel: Birkhäuser; 2009 [cité 27 févr 2020]. p. 239-55. (Progress in Inflammation Research). Disponible sur: https://doi.org/10.1007/978-3-7643-8903-1_13

71. Marcotte EL, Ritz B, Cockburn M, Yu F, Heck JE. Exposure to infections and risk of leukemia in young children. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. juill 2014;23(7):1195-203.

72. Gradel KO, Kaerlev L. Antibiotic use from conception to diagnosis of child leukaemia as compared to the background population: A nested case-control study. Pediatr Blood Cancer. juill 2015;62(7):1155-61.

73. Rudant J, Orsi L, Menegaux F, Petit A, Baruchel A, Bertrand Y, et al. Childhood acute leukemia, early common infections, and allergy: The ESCALE Study. Am J Epidemiol. 1 nov 2010;172(9):1015-27.

74. Law GR. Host, family and community proxies for infections potentially associated with leukaemia. Radiat Prot Dosimetry. 2008;132(2):267-72.

75. Adams KM, Nelson JL. Microchimerism: an investigative frontier in autoimmunity and transplantation. JAMA. 3 mars 2004;291(9):1127-31.

76. Von Behren J, Spector LG, Mueller BA, Carozza SE, Chow EJ, Fox EE, et al. Birth order and risk of childhood cancer: a pooled analysis from five US States. Int J Cancer. 1 juin 2011;128(11):2709-16.

77. Crump C, Sundquist J, Sieh W, Winkleby MA, Sundquist K. Perinatal and familial risk factors for acute lymphoblastic leukemia in a Swedish national cohort. Cancer. 1 avr 2015;121(7):1040-7.

78. Paltiel O, Lemeshow S, Phillips GS, Tikellis G, Linet MS, Ponsonby A-L, et al. The association between birth order and childhood leukemia may be modified by paternal age and birth weight. Pooled results from the International Childhood Cancer Cohort Consortium (I4C). Int J Cancer. 01 2019;144(1):26-33.

79. Urhoj SK, Raaschou-Nielsen O, Hansen AV, Mortensen LH, Andersen PK, Nybo Andersen A-M. Advanced paternal age and childhood cancer in offspring: A nationwide register-based cohort study. Int J Cancer. 01 2017;140(11):2461-72.

80. Li S, Kim E, Wong EM, Joo J-HE, Nguyen TL, Stone J, et al. Twin birth changes DNA methylation of subsequent siblings. Sci Rep. 16 2017;7(1):8463.

81. Slack JMW. Conrad Hal Waddington: the last Renaissance biologist? Nat Rev Genet. 2002;3(11):889-95.

82. Holliday R. The inheritance of epigenetic defects. Science. 9 oct 1987;238(4824):163-70.

83. Wu Ct null, Morris JR. Genes, genetics, and epigenetics: a correspondence. Science. 10 août 2001;293(5532):1103-5.

84. Rodenhiser D, Mann M. Epigenetics and human disease: translating basic biology into clinical applications. CMAJ Can Med Assoc J. 31 janv 2006;174(3):341-8.

85. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease. Nature. 2019;571(7766):489-99.

86. Rodríguez-Paredes M, Esteller M. Cancer epigenetics reaches mainstream oncology. Nat Med. mars 2011;17(3):330-9.

87. Hanly DJ, Esteller M, Berdasco M. Interplay between long non-coding RNAs and epigenetic machinery: emerging targets in cancer? Philos Trans R Soc Lond B Biol Sci. 05 2018;373(1748).

88. Vaissière T, Sawan C, Herceg Z. Epigenetic interplay between histone modifications and DNA methylation in gene silencing. Mutat Res. août 2008;659(1-2):40-8.

89. Rose NR, Klose RJ. Understanding the relationship between DNA methylation and histone lysine methylation. Biochim Biophys Acta. déc 2014;1839(12):1362-72.

90. Sawan C, Vaissière T, Murr R, Herceg Z. Epigenetic drivers and genetic passengers on the road to cancer. Mutat Res Mol Mech Mutagen. 3 juill 2008;642(1):1-13.

91. Ho L, Crabtree GR. Chromatin remodelling during development. Nature. 28 janv 2010;463(7280):474-84.

92. Micucci JA, Sperry ED, Martin DM. Chromodomain Helicase DNA-Binding Proteins in Stem Cells and Human Developmental Diseases. Stem Cells Dev. 15 avr 2015;24(8):917-26.

93. Mirabella AC, Foster BM, Bartke T. Chromatin deregulation in disease. Chromosoma. mars 2016;125(1):75-93.

94. Gallinari P, Marco SD, Jones P, Pallaoro M, Steinkühler C. HDACs, histone deacetylation and gene transcription: from molecular biology to cancer therapeutics. Cell Res. mars 2007;17(3):195-211.

95. Heard E, Martienssen R. Transgenerational Epigenetic Inheritance: Myths and Mechanisms. Cell. 27 mars 2014;157:95-109.

96. Ehrlich M. DNA hypermethylation in disease: mechanisms and clinical relevance. Epigenetics. déc 2019;14(12):1141-63.

97. Bestor TH. The DNA methyltransferases of mammals. Hum Mol Genet. oct 2000;9(16):2395-402.

98. Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell. 29 oct 1999;99(3):247-57.

99. Ziller MJ, Müller F, Liao J, Zhang Y, Gu H, Bock C, et al. Genomic distribution and inter-sample variation of non-CpG methylation across human cell types. PLoS Genet. déc 2011;7(12):e1002389.

100. Kohli RM, Zhang Y. TET enzymes, TDG and the dynamics of DNA demethylation. Nature. 24 oct 2013;502(7472):472-9.

101. Irizarry RA, Ladd-Acosta C, Wen B, Wu Z, Montano C, Onyango P, et al. Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat Genet. févr 2009;41(2):178-86.

102. Long MD, Smiraglia DJ, Campbell MJ. The Genomic Impact of DNA CpG Methylation on Gene Expression; Relationships in Prostate Cancer. Biomolecules. 14 2017;7(1).

103. Esteller M. Epigenetics in cancer. N Engl J Med. 13 mars 2008;358(11):1148-59.

104. Jones PA. The DNA methylation paradox. Trends Genet TIG. janv 1999;15(1):34-7.

105. Jjingo D, Conley AB, Yi SV, Lunyak VV, Jordan IK. On the presence and role of human gene-body DNA methylation. Oncotarget. 9 mai 2012;3(4):462-74.

106. Lu S, Davies PJA. Regulation of the expression of the tissue transglutaminase gene by DNA methylation. Proc Natl Acad Sci. 29 avr 1997;94(9):4692-7.

107. Jones PL, Veenstra GJ, Wade PA, Vermaak D, Kass SU, Landsberger N, et al. Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nat Genet. juin 1998;19(2):187-91.

108. Jones PA, Takai D. The role of DNA methylation in mammalian epigenetics. Science. 10 août 2001;293(5532):1068-70.

109. Li E. Chromatin modification and epigenetic reprogramming in mammalian development. Nat Rev Genet. sept 2002;3(9):662-73.

110. Bird A. DNA methylation patterns and epigenetic memory. Genes Dev. 1 janv 2002;16(1):6-21.

111. Mattick JS, Amaral PP, Dinger ME, Mercer TR, Mehler MF. RNA regulation of epigenetic processes. BioEssays News Rev Mol Cell Dev Biol. janv 2009;31(1):51-9.

112. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 15 févr 2001;409(6822):860-921.

113. Zheng Y, Joyce BT, Liu L, Zhang Z, Kibbe WA, Zhang W, et al. Prediction of genome-wide DNA methylation in repetitive elements. Nucleic Acids Res. 6 sept 2017;45(15):8697-711.

114. Ghantous A, Hernández-Vargas H, Byrnes G, Dwyer T, Herceg Z. Characterising the epigenome as a key component of the fetal exposome in evaluating in utero exposures and childhood cancer risk. Mutagenesis. 26 févr 2015;30.

115. Goyal D, Limesand S, Goyal R. Epigenetic responses and the developmental origins of health and disease. J Endocrinol. 1 juill 2019;242:T105-19.

116. Barrett JR. Programming the Future: Epigenetics in the Context of DOHaD. Environ Health Perspect. avr 2017;125(4):A72.

117. Barouki R, Melén E, Herceg Z, Beckers J, Chen J, Karagas M, et al. Epigenetics as a mechanism linking developmental exposures to long-term toxicity. Environ Int. 2018;114:77-86.

118. Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis. Am J Hum Genet. 7 avr 2016;98(4):680-96.

119. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 10 déc 2013;14(10):3156.

120. Bohlin J, Håberg SE, Magnus P, Reese SE, Gjessing HK, Magnus MC, et al. Prediction of gestational age based on genome-wide differentially methylated regions. Genome Biol. 07 2016;17(1):207.

121. 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. 8 mai 2012;13:86.

122. Rahmani E, Shenhav L, Schweiger R, Yousefi P, Huen K, Eskenazi B, et al. Genome-wide methylation data mirror ancestry information. Epigenetics Chromatin. 2017;10:1.

123. Lee HJ, Hore TA, Reik W. Reprogramming the methylome: erasing memory and creating diversity. Cell Stem Cell. 5 juin 2014;14(6):710-9.

124. Smallwood SA, Kelsey G. De novo DNA methylation: a germ cell perspective. Trends Genet TIG. janv 2012;28(1):33-42.

125. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 15 2018;555(7696):321-7.

126. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 pediatric leukemias and solid tumors. Nature. 15 mars 2018;555(7696):371-6.

127. Bolouri H, Farrar JE, Triche T, Ries RE, Lim EL, Alonzo TA, et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. Nat Med. 2018;24(1):103-12.

128. Panditharatna E, Filbin MG. The growing role of epigenetics in childhood cancers. Curr Opin Pediatr. févr 2020;32(1):67-75.

129. Feinberg AP. Phenotypic plasticity and the epigenetics of human disease. Nature. 24 mai 2007;447(7143):433-40.

130. Feil R, Fraga MF. Epigenetics and the environment: emerging patterns and implications. Nat Rev Genet. 4 janv 2012;13(2):97-109.

131. Krivtsov A, Figueroa M, Sinha A, Stubbs M, Feng Z, Valk P, et al. Cell of origin determines clinically relevant subtypes of MLL-rearranged AML. Leuk Off J Leuk Soc Am Leuk Res Fund UK. 13 déc 2012;27.

132. Feinberg AP, Koldobskiy MA, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet. mai 2016;17(5):284-99.

133. Huether R, Dong L, Chen X, Wu G, Parker M, Wei L, et al. The landscape of somatic mutations in epigenetic regulators across 1000 pediatric cancer genomes. Nat Commun. 8 avr 2014;5:3630.

134. Bender S, Tang Y, Lindroth AM, Hovestadt V, Jones DTW, Kool M, et al. Reduced H3K27me3 and DNA Hypomethylation Are Major Drivers of Gene Expression in K27M Mutant Pediatric High-Grade Gliomas. Cancer Cell. 11 nov 2013;24(5):660-72.

135. Schwartzentruber J, Korshunov A, Liu X-Y, Jones DTW, Pfaff E, Jacob K, et al. Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. Nature. 29 janv 2012;482(7384):226-31.

136. Ley TJ, Ding L, Walter MJ, McLellan MD, Lamprecht T, Larson DE, et al. DNMT3A mutations in acute myeloid leukemia. N Engl J Med. 16 déc 2010;363(25):2424-33.

137. Ghantous A, Hernandez-Vargas H, Herceg Z. DNA Methylation Analysis from Blood Spots: Increasing Yield and Quality for Genome-Wide and Locus-Specific Methylation Analysis. Methods Mol Biol Clifton NJ. 2018;1708:605-19.

138. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 15 sept 2014;23(R1):R89-98.

139. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. Clin Epigenetics. 27 août 2019;11(1):125.

140. Tikellis G, Dwyer T, Paltiel O, Phillips GS, Lemeshow S, Golding J, et al. The International Childhood Cancer Cohort Consortium (I4C): A research platform of prospective cohorts for studying the aetiology of childhood cancers. Paediatr Perinat Epidemiol. 2018;32(6):568-83.

141. Vineis P, Chadeau-Hyam M, Gmuender H, Gulliver J, Herceg Z, Kleinjans J, et al. The exposome in practice: Design of the EXPOsOMICS project. Int J Hyg Environ Health. 2017;220(2 Pt A):142-51.

142. Pruitt KD, Tatusova T, Maglott DR. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 1 janv 2005;33(suppl_1):D501-4.

143. 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. 1 oct 2011;98(4):288-95.

144. Perrier F, Novoloaca A, Ambatipudi S, Baglietto L, Ghantous A, Perduca V, et al. Identifying and correcting epigenetics measurements for systematic sources of variation. Clin Epigenetics. 2018;10:38.

145. 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. Bioinforma Oxf Engl. 15 mai 2014;30(10):1363-9.

146. Fortin J-P, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol. 3 déc 2014;15(11):503.

147. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. sept 2007;3(9):1724-35.

148. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS One. 2012;7(7):e41361.

149. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, L McKenney S, et al. DNA methylation of cord blood cell types: Applications for mixed cell birth studies. Epigenetics. 03 2016;11(5):354-62.

150. Benjamini Y, Hochberg Y. Controlling The False Discovery Rate - A Practical And Powerful Approach To Multiple Testing. J R Stat Soc Ser B. 30 nov 1995;57:289-300.

151. Bonferroni CE. Il calcolo delle assicurazioni su gruppi di teste. Studii in Onore del Profesor S. O. Carboni. Roma. 1936;

152. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinforma Oxf Engl. 1 sept 2010;26(17):2190-1.

153. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin. 2015;8:6.

154. Ambatipudi S, Cuenin C, Hernandez-Vargas H, Ghantous A, Le Calvez-Kelm F, Kaaks R, et al. Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study. Epigenomics. 11 févr 2016;8(5):599-618.

155. Woo HD, Fernandez-Jimenez N, Ghantous A, Degli Esposti D, Cuenin C, Cahais V, et al. Genome-wide profiling of normal gastric mucosa identifies Helicobacter pylori- and cancer-associated DNA methylome changes. Int J Cancer. 01 2018;143(3):597-609.

156. Chauvel C, Novoloaca A, Veyre P, Reynier F, Becker J. Evaluation of integrative clustering methods for the analysis of multi-omics data. Brief Bioinform. 14 févr 2019;

157. Merid SK, Novoloaca A, Sharp GC, Küpers LK, Kho AT, Roy R, et al. Epigenome-wide meta-analysis of blood DNA methylation in newborns and children identifies numerous loci related to gestational age. Genome Med. 2 mars 2020;12(1):25.

158. Alfano R, Plusquin M, Ghantous A, Keski-Rahkonen P, Chatzi L, Espín-Pérez A, Herceg Z, Kogevinas M, de Kok TM, Nawrot TS, Novoloaca A, Patel CJ, Pizzi C, Robinot N, Rusconi F, Scalbert A, Sunyer J, Vermeulen R, Vrijheid M, Vineis P, Robinson O, Chadeau-Hyam M. A Multi-Omic Analysis of Birthweight in

Newborn Cord Blood [Internet]. [cité 1 avr 2020]. Disponible sur: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3476782

159. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. déc 1986;51(6):1173-82.

160. Sobel ME. Asymptotic Confidence Intervals for Indirect Effects in Structural EQUATION MODELS. In: In Sociological Methodology. 1982. p. 290–312.

161. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinforma Oxf Engl. 15 2016;32(20):3150-4.

162. Zhang C-H. Nearly unbiased variable selection under minimax concave penalty. 25 févr 2010;38.

163. Mill J, Heijmans BT. From promises to practical strategies in epigenetic epidemiology. Nat Rev Genet. août 2013;14(8):585-94.

164. Breton Carrie V., Marsit Carmen J., Faustman Elaine, Nadeau Kari, Goodrich Jaclyn M., Dolinoy Dana C., et al. Small-Magnitude Effect Sizes in Epigenetic End Points are Important in Children's Environmental Health Studies: The Children's Environmental Health and Disease Prevention Research Center's Epigenetics Working Group. Environ Health Perspect. 1 avr 2017;125(4):511-26.

165. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. oct 2010;11(10):733-9.

166. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. Public Health Nutr. déc 2002;5(6B):1113-24.

167. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW. Cancer genome landscapes. Science. 29 mars 2013;339(6127):1546-58.

168. Simpkin AJ, Suderman M, Gaunt TR, Lyttleton O, McArdle WL, Ring SM, et al. Longitudinal analysis of DNA methylation associated with birth weight and gestational age. Hum Mol Genet. 1 juill 2015;24(13):3752-63.

169. Parets SE, Bedient CE, Menon R, Smith AK. Preterm Birth and Its Long-Term Effects: Methylation to Mechanisms. Biology. 21 août 2014;3(3):498-513.

170. Kwinta P, Pietrzyk JJ. Preterm birth and respiratory disease in later life. Expert Rev Respir Med. oct 2010;4(5):593-604.

171. Geldof CJA, van Wassenaer AG, de Kieviet JF, Kok JH, Oosterlaan J. Visual perception and visual-motor integration in very preterm and/or very low birth weight children: a meta-analysis. Res Dev Disabil. avr 2012;33(2):726-36.

172. Aarnoudse-Moens CSH, Weisglas-Kuperus N, van Goudoever JB, Oosterlaan J. Meta-analysis of neurobehavioral outcomes in very preterm and/or very low birth weight children. Pediatrics. août 2009;124(2):717-28.

173. Kerkhof GF, Breukhoven PE, Leunissen RWJ, Willemsen RH, Hokken-Koelega ACS. Does preterm birth influence cardiovascular risk in early adulthood? J Pediatr. sept 2012;161(3):390-396.e1.

174. Hille ETM, Weisglas-Kuperus N, van Goudoever JB, Jacobusse GW, Ens-Dokkum MH, de Groot L, et al. Functional outcomes and participation in young adulthood for very preterm and very low birth weight infants: the Dutch Project on Preterm and Small for Gestational Age Infants at 19 years of age. Pediatrics. sept 2007;120(3):e587-595.

175. Crump C, Sundquist K, Winkleby MA, Sieh W, Sundquist J. Gestational age at birth and risk of testicular cancer. Int J Cancer. 15 juill 2012;131(2):446-51.

176. Lu Y, Day FR, Gustafsson S, Buchkovich ML, Na J, Bataille V, et al. New loci for body fat percentage reveal link between adiposity and cardiometabolic disease risk. Nat Commun. 1 févr 2016;7:10495.

177. Huang X, Zhang H, Guo X, Zhu Z, Cai H, Kong X. Insulin-like growth factor 2 mRNA-binding protein 1 (IGF2BP1) in cancer. J Hematol OncolJ Hematol Oncol. 28 2018;11(1):88.

178. Mahaira LG, Katsara O, Pappou E, Iliopoulou EG, Fortis S, Antsaklis A, et al. IGF2BP1 expression in human mesenchymal stem cells significantly affects their proliferation and is under the epigenetic control of TET1/2 demethylases. Stem Cells Dev. 15 oct 2014;23(20):2501-12.

179. Parets S, Conneely K, Kilaru V, Fortunato S, Syed T, Saade G, et al. Fetal DNA Methylation Associates with Early Spontaneous Preterm Birth and Gestational Age. PloS One. 27 juin 2013;8:e67489.

180. Cuajungco MP, Podevin W, Valluri VK, Bui Q, Nguyen VH, Taylor K. Abnormal accumulation of human transmembrane (TMEM)-176A and 176B proteins is associated with cancer pathology. Acta Histochem. nov 2012;114(7):705-12.

181. Spiers H, Hannon E, Schalkwyk LC, Smith R, Wong CCY, O'Donovan MC, et al. Methylomic trajectories across human fetal brain development. Genome Res. mars 2015;25(3):338-52.

182. Yousefi P, Huen K, Davé V, Barcellos L, Eskenazi B, Holland N. Sex differences in DNA methylation assessed by 450 K BeadChip in newborns. BMC Genomics. 9 nov 2015;16:911.

183. Gruzieva O, Xu C-J, Yousefi P, Relton C, Merid SK, Breton CV, et al. Prenatal Particulate Air Pollution and DNA Methylation in Newborns: An Epigenome-Wide Meta-Analysis. Environ Health Perspect. 2019;127(5):57012.