The effect of genetic background and exposome on DNA methylation and its influence on human traits

PhD Candidate: Natàlia Carreras Gallo

TESI DOCTORAL UPF / 2022

Director de la tesi:

Dr. Juan Ramón González Ruiz

DEPARTMENT OF MEDICINE AND LIFE SCIENCES

PhD Programme in Biomedicine

Universitat Pompeu Fabra Barcelona

ISGlobal Barcelona Institute for Global Health

Als meus pares,
per estar sempre.

Al Pau,
per fer-ho tot fàcil.

Agradecimientos

Es difícil empezar a escribir cuando hay tantas personas que han formado parte de este proceso y sin las cuales no hubiese llegado donde estoy ahora.

Empezaré por Juan Ramón. Él ha sido mi supervisor de la tesis y la verdad es que no puedo estar más agradecida. Pocos son los supervisores que confían al 100% en sus estudiantes de doctorado, ofreciéndoles la máxima flexibilidad y adaptándose en todo momento a sus motivaciones. Muchas gracias por luchar en todo momento para que me sintiese a gusto con lo que estaba haciendo.

Otro pilar importante de mi tesis ha sido Alejandro… ¡Cuántas veces te he dicho que eres mi ángel de ISGlobal! Desde el primer momento he podido contar con tu ayuda sin pedir nada a cambio. Hemos hablado tanto de ciencia como de nuestras vidas y te has asegurado de que seguía motivada con el camino que estaba siguiendo. Sin ti, ahora no estaría aquí, y no sabes cuánto te lo agradezco. Espero que sigamos en contacto, ya que más que un compañero de trabajo, me llevo un buen amigo.

Ahora es el turno de Laura, a la que algunos conocerán como mi alma gemela. Esto lo explica que hayamos vivido mano a mano los últimos 7 años (misma carrera, máster, doctorado y compañeras de piso). Gràcies Laura, per tot. Per sentir-te sempre a prop, per fer-ho tot extremadament fàcil i per fer que sempre siguin bona idea les nostres bogeries.

A lo largo del doctorado, el BRGE ha ido creciendo y he podido conocer a grandes personas. Me llevo conmigo las paellas y las pizzas de los viernes que ayudaban a desconectar del trabajo por unas horas.

Luego está Carlos, el catalán pamplonica que me ha salvado de varios momentos de frustración. Gracias por echarme una mano siempre que lo he

necesitado. Qué suerte coincidir en el Congreso en Viena, fue un viaje para recordar.

No conozco a nadie que haya pasado por ISGlobal y no haya escuchado la pregunta: "¿Qué tal las vacaciones? ¿Bien o en familia?". Gracias Jose, por sacarnos una sonrisa todos los días, te echaré de menos.

Cómo olvidarme del equipo de *volley* que nos dio un primer puesto en el torneo del PRBB. Gracias infinitas a todos por crear un ambiente tan acogedor en el equipo y por motivarnos a mejorar en cada partido. Gracias a los predocs, postdocs y estadísticos que han pasado por ISGlobal, por las mil charlas a la hora de comer. En especial, gracias a mi compañera de PhD representatives, Eve, por hacerme reír siempre.

Me gustaría dar un especial agradecimiento a Luis, por ser mi tutor y ayudarme a contextualizar los resultados en un marco genético. A Gemma Punyet, por cuidarnos como sus hijas de ISGlobal. And special thanks to Varun and Ryan, for giving me the opportunity to collaborate with the TruDiagnostic company. Gracias también a los proyectos y a las becas que han financiado parte o la totalidad de mi investigación, en particular a las ayudas ofrecidas en el Pla estratègic de recerca i innnovació en Salut (PERIS) de la Generalitat de Catalunya.

Por último, pero no menos importante, gracias a toda mi familia por apoyarme siempre. A mis padres, por crear la mejor familia que una podría pedir. A mi hermano, por preocuparse por mí y ser mi referente. A mi abuela, por quererme incondicionalmente. A Pau, por la super portada de tesis y por hacerme extremadamente feliz y solo imaginarme una vida a su lado.

Barcelona, 2 de noviembre de 2022

Natàlia Carreras

“Above all, don’t fear difficult moments.
The best comes from them.”

Rita Levi-Montalcini (1909-2012)

Abstract

DNA methylation is a biological process defined as the addition of a methyl group into a DNA molecule. Among others, genetic and environmental factors are the main modulators of this process. DNA methylation alterations are associated with many common diseases. Thus, it is extremely important to identify which factors are leading to methylation changes and, consequently, to disease development.

In this thesis, we have evaluated the effect of specific genetic variants and environmental factors, as well as their interaction, in DNA methylation. To this end, we used two main study populations: Human Early-Life Exposome (HELIX) and the TruDiagnostic DNA Biobank.

First, we detected genome-wide differentially methylated CpG sites for tobacco, alcohol, and marijuana consumption. Importantly, the top alcohol-related CpG site mediated 73.6% of the effect of alcohol consumption on hypertension. Second, we found genotype-specific methylation patterns in three common polymorphic inversions (8p23.1, 16p11.2, and 17q21.31). Additionally, we identified multiple significant inversion-exposure interactions. Finally, we identified a prenatal environment, defined as the combination of four environmental exposures, where boys were more likely of being obese than girls. We designed an algorithm to predict this environment using the methylome.

Our findings suggest new genetic and environmental factors modulating DNA methylation that should be considered for new targets in disease prevention. In addition, personalized medicine is now on the horizon since the effect of environmental factors on DNA methylation depends on the personal genetic background, as well as the combination of multiple exposures.

Resum

La metilació de l'ADN és un procés biològic que es defineix com l'adició d'un grup metil a la molècula d'ADN. Entre altres, els factors genètics i ambientals són els principals moduladors d'aquest procés. Les alteracions en la metilació s'associen amb nombroses malalties comuns. Per això, és molt important identificar quins factors donen lloc a canvis en la metilació i, per conseqüència, en el desenvolupament de malalties.

En aquesta tesis, hem avaluat l'efecte de variants genètiques i factors ambientals, així com la seva interacció, en la metilació de l'ADN. Amb aquesta fi, hem utilitzat dues cohorts*: Human Early-Life Exposome* (HELIX) i el Biobanc d'ADN de TruDiagnostic.

En primer lloc, hem detectat llocs CpG diferencialment metilats segons el consum de tabac, alcohol i marihuana. Més important encara, el 73.6% de l'associació entre el consum d'alcohol i la hipertensió estava mediatitzat pel CpG més associat amb alcohol. En segon lloc, hem trobat patrons de metilació específics del genotip de tres inversions polimòrfiques (8p23.1, 16p11.2 i 17q21.31). Addicionalment, hem identificat múltiples interaccions inversió-exposició significatives entre aquestes inversions i exposicions ambientals. Per últim, hem identificat un ambient prenatal, definit com la combinació de quatre exposicions, on els nois tendeixen a ser més obesos que les noies. A més a més, hem dissenyat un algoritme per predir aquest ambient utilitzant la metilació.

Els nostres resultats suggereixen nous factors genètics i ambientals moduladors de la metilació que haurien de ser considerats en la prevenció de malalties. Així mateix, s'obre camí a la medicina personalitzada, ja que l'efecte dels factors ambientals depèn del context genètic, així com de la combinació de vàries exposicions.

Resumen

La metilación del ADN es un proceso biológico definido como la adición de un grupo metilo a la molécula de ADN. Entre otros, los factores genéticos y ambientales son los principales moduladores de este proceso. Las alteraciones en la metilación se asocian con enfermedades comunes. Por ello, es crucial identificar qué factores producen cambios en la metilación y, por consiguiente, el desarrollo de enfermedades.

En esta tesis, hemos evaluado el efecto de variantes genéticas y factores ambientales, así como su interacción, en la metilación del ADN. Para ello, hemos usado dos cohortes: Human Early-Life Exposome (HELIX) y el Biobanco de ADN de TruDiagnostic.

En primer lugar, hemos detectado sitios CpG diferencialmente metilados según el consumo de tabaco, alcohol y marihuana. Notablemente, el 73.6% de la asociación entre el consumo de alcohol y la hipertensión estaba mediado por el CpG más asociado a alcohol. En segundo lugar, hemos encontrado patrones de metilación específicos del genotipo de tres inversiones polimórficas (8p23.1, 16p11.2 i 17q21.31). Además, hemos identificado múltiples interacciones inversión-exposición significativas. Por último, hemos identificado un ambiente prenatal, definido como la combinación de cuatro exposiciones, donde los niños son más obesos que las niñas. Hemos diseñado un algoritmo para predecir este ambiente usando la metilación.

Nuestros resultados sugieren nuevos factores genéticos y ambientales moduladores de la metilación que podrían ser claves en la prevención de enfermedades. Asimismo, se abre camino a la medicina personalizada, ya que los factores ambientales dependen del contexto genético, así como de la combinación de varias exposiciones.

Preface

This thesis was written at the Barcelona Institute for Global Health (ISGlobal) within the Bioinformatic Research Group in Epidemiology (BRGE). It was supervised by Dr. Juan Ramón González.

This work consists of a compilation of 3 scientific publications (1 published, 1 under review, and 1 submitted) co-authored by the PhD candidate. This agrees with the procedures of the PhD program in Biomedicine, organized by the Department of Medicine and Life Sciences of the Universitat Pompeu Fabra (UPF).

The present thesis contributed to: (1) identify new epigenetic targets in the association between drug consumption and common diseases, such as hypertension neurological disorders; (2) identify methylomic patterns associated with common inversion genotypes; (3) propose interactions between inversions and environmental exposures that need further validation; (4) recognize an environment where girls are likely more protected from obesity and neurological delay than boys; (5) discuss future research directions to study methylation changes in personalized medicine.

Abbreviations

450K: Infinium HumanMethylation450 BeadChip

5caC: 5-carboxylcytosine

5fC: 5-formylcytosine (5fC)

5hmC: 5-hydroxymethylcytosine

5mC: 5-methylcytosine

ANT: Attention Network Test

BER: Base Excision Repair

BMI: Body Mass Index

bp: base-pair

cfDNA: cell-free DNA

circRNA: circular RNA

CNV: Copy Number Variant

CpG: 5'-Cytosine-phosphate-Guanine-3

CVD: Cardiovascular Disease

DMP: Differentially Methylated Positions

DMR: Differentially Methylated Regions

DNA: Deoxyribonucleic Acid

DNMT: DNA Methyltransferase

EPIC: HumanMethylationEPIC BeadChip

eQTL: Expression Quantitative Trait Loci

EWAS: Epigenome Wide Association Study

GO: Gene Ontology

GxE: Gene-environment interaction

GWAS: Genome Wide Association Study

HELIX: Human Early Life Exposome

KEGG: Kyoto Encyclopedia of Genes and Genomes

mQTL: Methylation Quantitative Trait Loci

mRNA: messenger RNA

miRNA: microRNA

ncRNA: non-coding RNA

OC: Organochlorine

OP pesticides: Organophosphate Pesticides

OR: Odd Ratio

PCA: Principal Component Analysis

PERS: Polyenviromic or Poyenvironmental Risk Score

piRNA: piwi-interacting RNA

PBDE: Polybrominated Siphenyl Ether compounds

PFAS: Perfluorinated Alkylated substances

PM2.5: Particulate Matters with a diameter of 2.5µm

PM10: Particulate Matters with a diameter of 10µm

PRS: Polygenic Risk Score

RE: Repetitive Element

RNA: Ribonucleic Acid

siRNA: small interfering RNA

snRNA: small nuclear RNA

snRNP: small nuclear ribonucleoprotein

SNP: Single Nucleotide Polymorphism

SV: Structural Variant

SVA: Surrogate Variable Analysis

TABLE OF CONTENT

1 | INTRODUCTION

1.1. Epigenetics overview

Genetics is widely known and studied, and it refers to the alteration of gene activation and function due to changes in DNA sequence. Conversely, epigenetics is defined as the study of inherited changes in gene activity or function without modifying the DNA sequence. Even though genetic information is the same in all cells of an organism, not all genes are expressed in all cell types [1]. Epigenetics is responsible for creating tissue-specific gene expression profiles. Therefore, epigenetic processes are natural and essential to many organism functions. However, if epigenetic modifications occur improperly, there can be major adverse health and behavioral effects [2,3].

One of the main factors altering epigenetics is exposure to environmental hazards, like tobacco smoke or air pollution [4,5]. In this way, epigenetics plays a bridge role between our genes and our behaviors and environment. Additionally, epigenetic changes may persist through multiple cell divisions and may also last for multiple generations [6–9]. The main epigenetic mechanisms include DNA methylation, histone modifications, and non-coding RNAs.

1.1.1. DNA methylation

DNA methylation is the most known epigenetic process, and it is defined as the addition of a methyl group (CH_3) to the carbon-5 position of cytosine in DNA, resulting in a 5-methylcytosine (5mC) (**Figure 1**) [10,11]. DNA methylation is responsible for recruiting proteins involved in gene repression, increasing or reducing gene expression [1].

Since DNA methylation is the main topic of this thesis, the next sections include more information about this epigenetic mark.

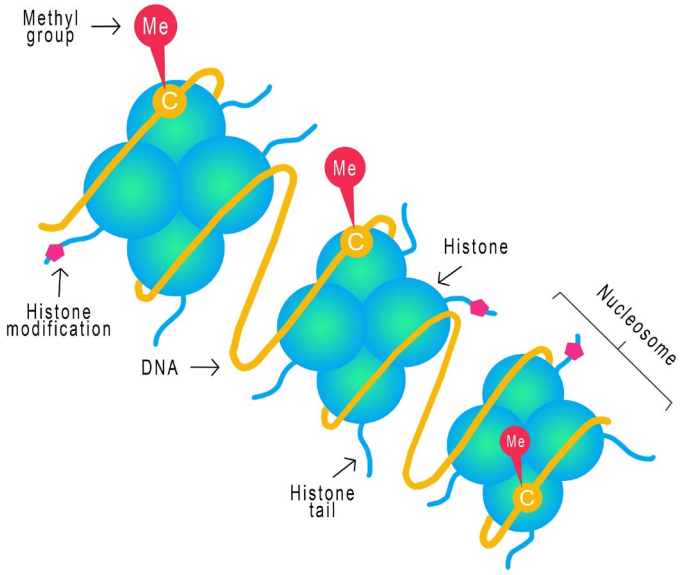


Figure 1 | Chromatin structure. DNA (orange chain) is packaged around histone octamers (blue balls) that constitute nucleosomes. DNA methylation (red balls) occurs in position 5 of cytosines (orange balls). Histone modifications are represented as pink pentagons. Based on Epigentek (www.epigentek.com).

1.1.2. Histone modifications

DNA is compacted to fit into the cell nucleus as chromatin. Histones are the proteins that act in packaging the DNA double-helix into structural units called nucleosomes. These units are octamers of four core histones (H2A, H2B, H3, and H4) that wrap 147 base pairs of DNA in two turns (**Figure 1**). Moreover, 50 base pairs separate one nucleosome from the next one [12,13]. Importantly, an extra histone (linker histone H1) is

required to further condense the chromatin and form the higher order structures (**Figure 2**) [14].

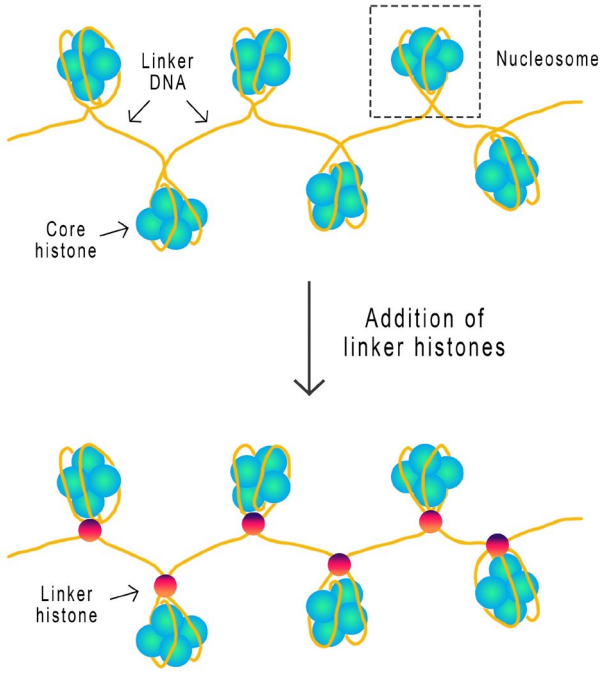


Figure 2 | Chromatin structures condensed by linker histones. Based on Zhou et al. [14]**.**

Although the main function of nucleosomes is packaging DNA, they also ensure or impair the DNA sequence's accessibility to proteins involved in DNA replication, recombination, gene expression, and DNA repair [13]. Therefore, chromatin can be condensed and decondensed due to histone modifications depending on the cell's needs [15]. Heterochromatin is known as the type of chromatin that is tightly packed and condensed. Heterochromatin is typical of inactive genes since the

condensed structure inhibits the recruitment of RNA polymerase and gene regulatory proteins. On the other hand, euchromatin is defined as loosely wrapped chromatin. In this case, DNA is more accessible and, therefore, it is actively transcribed [16].

Histone modifications occur at histone tails by adding acetyl, methyl, phosphate, ubiquitin, or other groups [16,17]. The histone modification can impact gene expression by altering chromatin structure or recruiting histone modifiers [17]. Generally, histone acetylation is associated with highly transcriptionally regions, whereas hypoacetylated histones are found in transcriptionally inactive regions [18]. Histone methylation can be found in expressed and non-expressed regions according to the specific position within the histone and within the gene.

1.1.3. Non-coding RNAs

The group of RNAs that do not encode functional proteins is called non-coding RNAs (ncRNAs). According to their regulatory roles, ncRNAs can be classified into two categories (**Table 1**). Housekeeping ncRNAs are abundant and widely expressed in cells, regulating primary cellular functions, such as translating messenger RNA (mRNA) into proteins [19]. Regulatory ncRNAs encompass many types of RNAs and play important roles in gene expression regulation at epigenetic, transcriptional, and post-transcriptional levels [20].

Among other functions, these regulatory non-coding RNAs can induce DNA methylation and histone modification, the other two major groups of epigenetics [21]. For instance, microRNAs (miRNAs) and small interfering RNAs (siRNAs) are involved in the RNA interference

pathway, which represses translation by neutralizing target complementary transcripts (**Figure 3**). Piwi-interacting RNAs (piRNAs) are involved in transposon silencing and other epigenetic pathways. Small nuclear RNAs (snRNAs) have the role to remove the introns of a precursor mRNA by establishing small nuclear ribonucleoprotein complexes (snRNPs). These snRNPs are combined to form a large ribonucleoprotein complex called spliceosome. Using this mechanism, a mature transcript can be created during the splicing process (**Figure 4**).

Table 1 | Types of non-coding RNAs (ncRNAs). Housekeeping and regulatory ncRNAs are the two main groups according to regulatory roles. Each division encompasses seven types of ncRNAs of different sizes. Table from Zhang [19].

Type	Abbreviation	Full name	Size
Housekeeping ncRNAs	rRNA	ribosomal RNA	120-4,500 nt
	tRNA	transfer RNA	76-90 nt
	snRNA	small nuclear RNA	100-300 nt
	snoRNA	small nucleolar RNA	60-400 nt
	TERC	telomerase RNA	/
	tRF	tRNA-Derived Fragments	16-28 nt
	tiRNA	tRNA halves	29-50 nt
Regulatory ncRNAs	miRNA	microRNA	21-23 nt
	siRNA	small interfering RNA	20-25 nt
	piRNA	Piwi-interacting RNA	26-32 nt
	eRNA	enhancer RNA	50-2,000 nt
	lncRNA	Long non-coding RNA	> 200 nt
	circRNA	circular RNA	100-10,000 nt
	Y RNA	Y RNA	/

Circular RNAs (circRNAs) are the unique class of ncRNAs that form covalently closed loop structures. They play important roles in regulating alternative RNA splicing or transcription, as well as acting as competing endogenous RNAs [19,22].

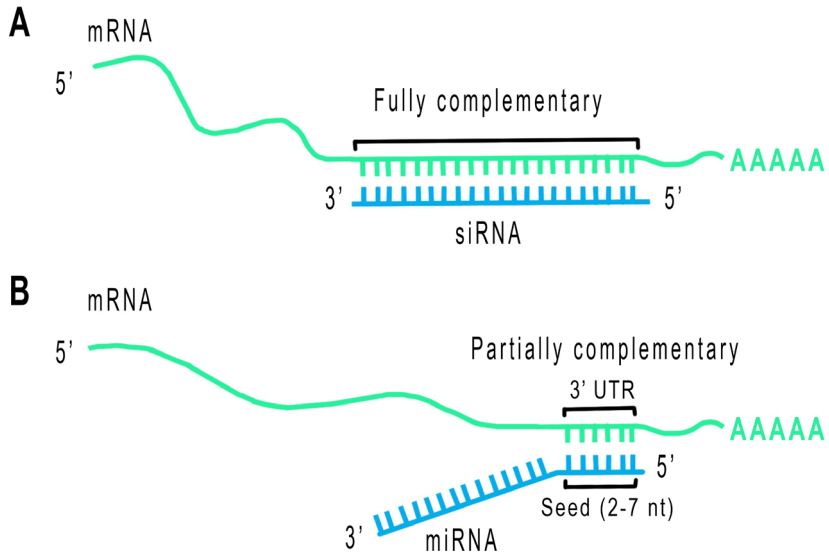


Figure 3 | Target recognition by small interfering (siRNA) and microRNA (miRNA). (A) siRNA is usually fully complementary to the coding region of its target mRNA; (B) miRNA is partially complementary to its target miRNA. Complementary binding usually occurs at the seed region (nucleotides (nt) 2–7 of the 5' end) of miRNA and the 3' UTR of the target mRNA. Based on Lam et al. [23].

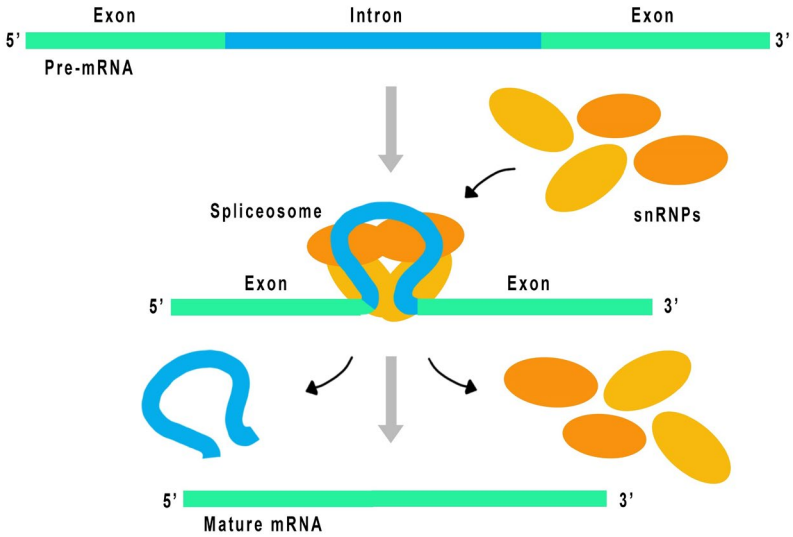


Figure 4 | Splicing process where a transcribed pre-mRNA is transformed into a mature mRNA. The pre-mRNA contains exons (green) and introns (blue). To remove introns, the small nuclear RNA (snRNAs) molecules bind to specific proteins to form a small nuclear ribonucleoprotein complex (snRNP). Multiple snRNPs (yellow and orange) are combined to form the spliceosome. The spliceosome removes the introns from the pre-mRNA leading to a mature mRNA.

1.2. DNA methylation overview

1.2.1. DNA methylation distribution

DNA methylation mostly occurs at CpG sites, defined as a cytosine followed by a guanine, separated by a phosphate group, in the same strain along its 5' → 3' direction. In short, CpG is the abbreviation for 5' – C – phosphate – G – 3' (**Figure 5**). In mammals, CpG sites are less frequent than would be expected from the base composition. In human DNA, the GC fraction is 0.4, thus, we would expect a frequency of 0.04 (0.2 x 0.2) to occur for CpG sites. However, this frequency is about 0.008. This is explained by the common mutation where 5mC deaminates to thymine [24].

Figure 5 | A CpG site. On the left, the green strand contains a CpG site from 5' to 3'. On the reverse DNA strand (blue), the complementary CpG site is shown. On the right, a C-G base-pairing is also indicated to distinguish it from a CpG. Based on Wikipedia (www.wikipedia.org).

Although a low proportion of CpG sites, these sites are highly methylated. In humans, embryonic stem cells present up to 80 % of the CpG sites methylated, except for the CpG islands, which tend to be unmethylated [25–27]. CpG islands are regions of 1 kb approximately with high CpG density near the gene promoters that are often not methylated. More than 70% of mammalian promoters reside within CpG

islands, especially within housekeeping and developmentally regulated gene promoters. They are unmethylated to avoid deamination and consequent mutation to thymine during evolution [1,11]. CpG islands have been evolutionarily conserved to promote gene expression since they promote the accessibility of DNA and enhance transcription factor binding [1].

Recent studies in humans have demonstrated that cytosine methylation also occurs when cytosines are not followed by a guanine. This is known as non-CG methylation (mCH) where H corresponds to adenine (A), cytosine (C), or thymine (T). Although methylation of CH dinucleotides has been found in pluripotent and brain cells, it is still unclear their function [28].

1.2.2. DNA methylation and demethylation processes

DNA methylation occurs when DNA methyltransferase (DNMT) enzymes transfer a methyl group from the cofactor S-Adenosyl-L methionine (SAM) to the 5-carbon in cytosine in DNA [29]. There are two main groups of DNMTs (**Figure 6**). The first one includes those enzymes that are responsible for de novo methylated sites (DNMT3a and DNMT3b). This means establishing DNA methylation marks at previously unmethylated sites and occurs during early development. Although the specific mechanism whereby de novo DNMTs target DNA regions is partially understood, there is evidence that histone tail modifications and ncRNAs are involved in guiding the establishment of 5mC [10,30,31]. The second group comprises DNMTs that aim to maintain already established methylated sites during DNA replication (DNMT1). During this process, the DNA chain is divided into two

strands. When the complementary sequences are synthesized, the new chains are hemimethylated because only the old strand is methylated. At this moment, DNMTs are recruited, and they methylate the appropriate cytosines following the same pattern as in the old chain (**Figure 6**) [10,32,33].

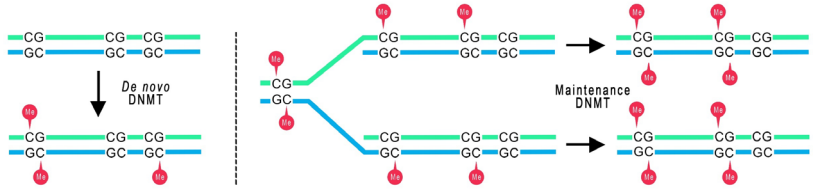


Figure 6 | *De novo* and maintenance DNA methyltransferases (DNMTs). On the left, *de novo* DNA methylation establishes methylated cytosines at previously unmethylated sites. On the right, maintenance DNA methylation involves recognition of hemi-methylated CpG sites generated during DNA replication and methylation of the newly synthesized strand following the same pattern as the old strand. Based on Schmitz et al. [10].

DNA demethylation occurs in many methylated cytosines along the genome, especially within gene bodies. This process consists of the conversion of a 5mC to an unmodified cytosine (C). DNA demethylation can occur passively or actively. Passive DNA demethylation happens when 5mC disappears from the genome due to a lack of maintenance DNMTs. Oppositely, active DNA demethylation involves the oxidation of methylated cytosines by ten-eleven translocation (TET) enzymes into oxidized derivatives of 5mC [34].

Active DNA demethylation takes place in a cycle, which starts at 5mC and finishes with an unmodified cytosine (**Figure 7**). First, 5mC is

oxidized to 5-hydroxymethylcytosine (5hmC), which in turn is oxidized to 5-formylcytosine (5fC), and finally is oxidized to 5-carboxylcytosine (5caC). Among these molecules, 5fC and 5caC can be converted to unmodified cytosines by thymine DNA glycosylase (TDG) in combination with base excision repair (BER). Moreover, 5hmC is the most stable oxidized derivative of 5mC. These molecules have been of great interest to many recent studies, which suggest a potential role in stable epigenetic roles [34].

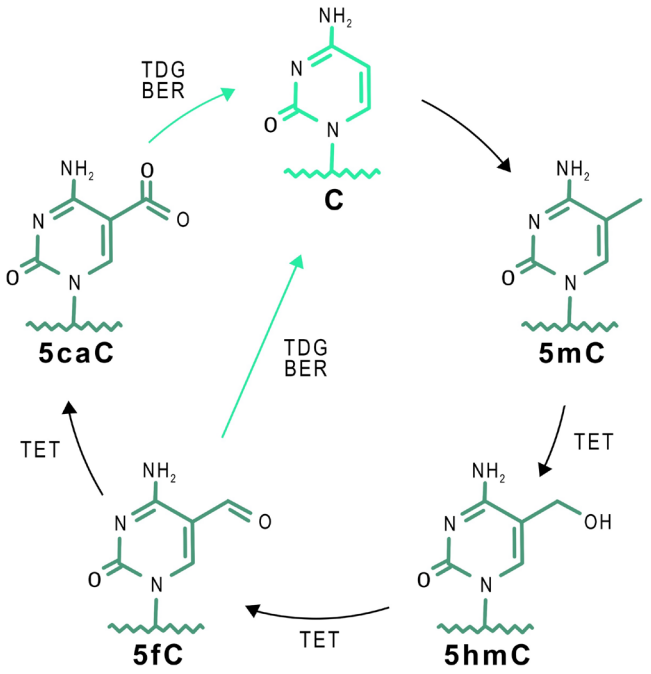


Figure 7 | Active DNA demethylation cycle. Cytosines (C) can be methylated through DNA methyltransferase (DNMT) enzymes, resulting in 5-methylcytosine (5mC). 5mC is oxidized to 5-hydroxymethylcytosine (5hmC) by ten-eleven translocation (TET) enzymes, which is oxidized to 5-formylcytosine (5fC), which in turn is oxidized to 5-carboxylcytosine (5caC). Thymine DNA glycosylase (TDG) and base excision repair (BER) convert 5fC and 5caC into C. Based on abcam [35].

1.2.3. Functions of DNA methylation

DNA methylation is crucial for cell differentiation and development, being essential for cell viability. Experiments removing DNMT genes in mouse models were incompatible with the viability of the organism [30,36].

Although CpG islands within promoters tend to be unmethylated, environmental exposures and genetic factors can produce changes at methylation levels [37–40]. In general, methylation at promoters leads to gene transcription repression. The effect of methylation on gene expression is not direct, but methylation can modulate the chromatin structure and, consequently, reduce the accessibility of the binding factors [11,41]. Even though this is a way of silencing, the inactivation of CpG island promoters through histone methylation is also common because is more plastic than DNA methylation [42]. Nevertheless, there are three major groups of genes whose silencing is stable and is regulated by DNA methylation: genes on the inactive X chromosome, imprinted genes, and germline-specific genes [11].

Oppositely, gene bodies present high levels of methylation and are positively correlated with transcription [43]. This may be explained by two hypotheses. First, DNA methylation in the gene body facilitates transcription elongation and splicing. Therefore, it seems that while transcription initiation is sensible to DNA methylation silencing, elongation is not [30]. The role of methylation in splicing has been suggested because of the change of methylation level at exon-intron boundaries, being exons more methylated than introns [44]. Second, it avoids the transcription initiation at cryptic promoters [11]. Methylation

in a downstream promoter would avoid the transcription from that promoter but it would allow the elongation from the first promoter [45].

DNA methylation also provides genome stability by repressing the activation of repetitive elements (REs). These elements constitute more than a third of the genome and are dangerous because they can recombine and lead to chromosomal rearrangements. DNA methylation is responsible for compacting DNA where REs are present, protecting it from the recombination and mobility of these elements. Previous studies have demonstrated that hypomethylation in repetitive sequences is common in complex diseases, like cancer, where genome instability is relevant [11,46].

1.2.4. DNA methylation and health

Many of the diseases that cannot be explained by genetics, may be explained by epigenetics. For example, monozygotic twins with the same genetic variant for a disease can present very different clinical characteristics. In this case, genetics is not responsible for the phenotypic traits and the twins may have epigenetic differences that lead to these differences. In the same line, genetic and epigenetic mutations can induce the same phenotype. For instance, one gene can be downregulated due to a gene disruption (genetic mutation) or due to the high condensation of the chromatin, making impossible its transcription (epigenetic mutation). Finally, genetic and epigenetic mutations are not exclusive. Genetic alterations may affect proteins involved in the epigenetic processes, such as DNMTs, producing epigenetic alterations at multiple loci [3].

Epigenetic modifications are maintained during cell division due to the maintenance of DNMT enzymes. Therefore, in one organ, one cell will be divided into two cells with the same epigenome [3,10]. In addition, epigenetic changes can be transmitted to the next generations if they affect the germ line.

The environment is the third key element in the relationship between genetics and epigenetics. While genetics predispose to diseases that are highly influenced by the environment, environmental factors, such as diet or air pollution, alter the epigenome. Thus, identifying environmental factors altering epigenetics for specific diseases provides information for designing biomarkers for preventing disease [3].

Epigenetics is involved in the regulation of all biological processes in the body from conception to death and it regulates development and adaptations during the lifetime. Therefore, epigenetic alterations may result in disorders [2]. Many studies have demonstrated the important role of DNA methylation in common diseases, such as cancer, autoimmune diseases, neurological disorders, cardiovascular disease (CVD), and obesity [2]. Besides, recent studies also highlight the importance of methylation alterations in rare diseases [47].

- **Cancer**

In patients with cancer, it is common to find hypermethylation of promoter regions of tumor suppressor genes, leading to an inactivation of tumor suppressor functions. On the other hand, global hypomethylation has been found in cancer cells, which contributes to genomic instability and cell transformation. Additionally, DNA

methylation has an important role in regulating cell division, DNA repair, cell differentiation, apoptosis, angiogenesis, metastasis, growth factor response, detoxification, and drug resistance [48]. All these processes are highly involved in tumor evolution and specific methylation marks for these disturbances can be used to early detect cancer. It is worth mentioning that not only DNA methylation is being used as a cancer biomarker, but also other epigenetic factors, like miRNAs [49].

Importantly, DNA methylation is reversible, so these marks are extremely important for therapeutic approaches. Another benefit of studying epigenetic changes in cancer versus genetic changes is that some methylation abnormalities involved in tumorigenesis appear before mutations. Thus, detecting abnormal epigenetic marks may help to detect cancer at the very beginning stages. Besides, many methylation changes are tissue-specific, such as the hypermethylation at *BRCA1* exclusively in breast and ovarian carcinomas [50].

One strategy in current clinical practice for cancer treatment is the use of DNMT inhibitors. Although global hypomethylation may contribute to tumor progression, the inhibition of DNMTs helps to restore the activity of tumor suppressors and genes involved in important cellular pathways [51]. Various studies confirm tumorigenesis inhibition after drug intake. In particular, Vidaza and Decitabine are two anticancer drugs approved by the Food and Drug Administration (FDA) that are currently used for the treatment of myelodysplastic syndromes (MDS) and chronic myelomonocytic leukemia (CML) [52,53].

Even more interestingly, the study of DNA methylation in liquid biopsies and, in particular, the analysis of cell-free DNA (cfDNA) is a

potential non-invasive diagnostic approach in oncology [54]. cfDNA is composed of fragments of DNA released by cells into the circulation, mainly due to cell death. In cancer patients, cfDNA also contains DNA from tumor cells. Therefore, the detection and analysis of tumor-derived cfDNA is a potential strategy for evaluating tumor diagnosis, progression, or treatment rejection. As an example, 7 cfDNA methylation markers in plasma were able to discriminate epithelial ovarian cancer from benign pelvic masses with a sensitivity of 90.6% at a specificity of 89.7% [55].

- **Autoimmune diseases**

Many studies suggest that autoimmune diseases are also regulated by epigenetic factors. Indeed, hematopoietic lineage, antigen-receptor, allelic exclusion, and inducible immune response against pathogens are epigenetically regulated [56]. Rheumatoid arthritis, a common autoimmune disease, is associated with hypermethylation of *HDAC1* and *HDAC2* genes [57]. These genes encode two histone deacetylase proteins, leading to histone tail modifications. Besides, patients with multiple sclerosis present lower methylation levels in central nervous system white matter compared to controls [58].

- **Neurological disorders**

Epigenetics is particularly important in brain development and disease. Day and Sweatt introduced the discipline ‘Neuroepigenetics’ in 2010, defined as the study of the mechanisms that allow dynamic regulation of the epigenome in nondividing cells of the nervous system, along with

the epigenetic process involved in neuronal differentiation and cell-fate determination [59]. During these years, many studies have demonstrated the important role of epigenetic alterations in learning, memory, and synaptic plasticity, as well as in neurodevelopmental and neuropsychiatric disorders, neurodegeneration, and aging [60]. As an example, the emergence of Alzheimer's disease and schizophrenia has been seen associated with global hypomethylation due to low levels of DNMT3a mRNA and lower methylation levels at REs [61].

- **Cardiovascular disease (CVD)**

The role of epigenetics in CVD is complex and partially known. There are many risk factors for these diseases that modulate the epigenome, such as sex, smoking, aging, and diabetes [62]. Hypertension is a major risk factor for stroke and CVD and its pathogenesis seems to be modulated by DNA methylation. For example, global hypomethylation in peripheral blood leucocytes has been seen in hypertensive patients. Moreover, those patients have high methylation levels at the *ACE* gene promoter, inactivating the catalyzation of the angiotensin II production crucial for blood pressure regulation [63]. Other studies revealed lower methylation levels in hypertensive patients in two CpG sites that mapped to two genes *SLC7A11* and *PHGDH*, respectively [64–66]. While *SLC7A11* was associated with vascular tone regulation, *PHGDH* regulates serin synthesis and tissue growth [66].

- **Obesity**

Obesity is defined as abnormal or excessive fat accumulation. People with obesity have more risk of suffering from type 2 diabetes, CVD, hypertension, or metabolic syndrome, among others. Several factors are known to affect abnormal fat accumulation, such as lifestyle behaviors (dietary habits, physical exercise, and sleep patterns), social factors (educational level and economic status), endocrine disorders (hypothyroidism), or intake of certain drugs (like corticosteroids). However, there is high inter-individual variability when evaluating the effect of these factors. This suggests that genetic and epigenetic elements may have an important role in the pathology of obesity [67].

Many studies have demonstrated the significant association between obesity and changes in DNA methylation levels in many tissues. For example, Aslibekyan et al. compared obese with normal-weight adults and reported methylation changes in the carnitine palmitoyltransferase 1A (*CPT1A*) gene [68]. This gene encodes for a protein implicated in the control of fasting triglycerides and lipoprotein levels. Another study revealed several CpG sites differentially methylated in cord blood when comparing newborns from obese mothers with newborns from normal-weight mothers [69].

- **Rare diseases**

Rare diseases are defined as disorders with a prevalence of less than 1 person in 2,000 people. Among the 7,000 different rare diseases described worldwide, 80% are thought to have a genetic origin [70]. Moreover, about 60% of the patients remain undiagnosed [71]. The high

number of undiagnosed patients may be explained, in part, due to a non-genetic origin of the disease.

Epimutations are rare alterations in DNA methylation patterns. They are identified by detecting groups of contiguous CpG sites with aberrant methylation values compared with the reference population. Previous studies demonstrated that epimutations could explain the development of rare diseases [47,72,73]. For instance, Aref-Eshghi et al. diagnosed 67 individuals with uncertain clinical diagnoses related to neurodevelopmental presentations and congenital anomalies by identifying specific epimutations [72].

1.2.5. DNA methylation assessment

Many technologies and strategies have been developed to measure and analyze DNA methylation. According to the biological question, it would be better to use one or another methodology. They can be classified according to three key factors: DNA methylation extends, pretreatment, and analytical step.

- **DNA methylation extends**

DNA methylation can be estimated at the global level or DNA methylation sites. While global-level strategies assess overall changes, the second strategy evaluates specific alterations in DNA methylation. Additionally, specific changes at DNA methylation sites can be estimated at particular genes or in the whole genome. Epigenome-wide association studies (EWAS) examine genome-wide epigenetic variants, mainly DNA methylation levels at CpG sites, to detect significant

differences associated with phenotypes of interest [74]. These studies have helped to understand the role of methylation in many diseases. Since the number of CpG sites evaluated is very large (can reach 850,000 CpG sites), bioinformatic tools to process, analyze, and interpret methylation data from EWAS have evolved. Typically, results from EWAS are displayed in a Manhattan plot, where each point represents a CpG site along the chromosomes and the Y axis reflects the level of significance (**Figure 8**).

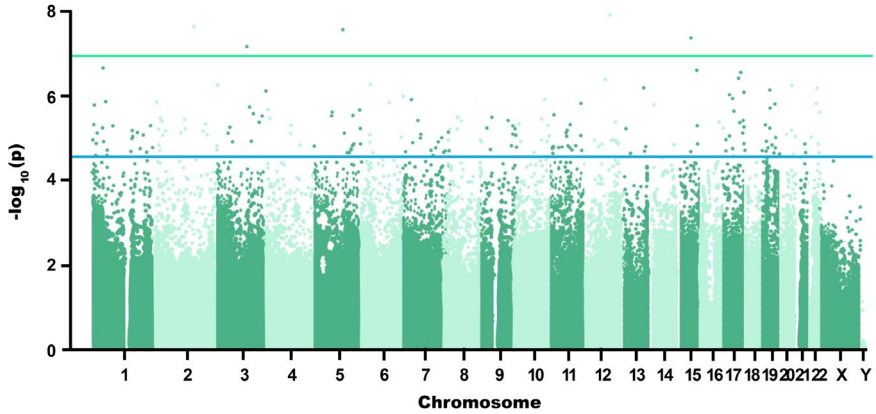


Figure 8 | Example of a Manhattan plot for EWAS analysis. The x-axis shows the genome position of the CpG sites. The y-axis shows the $-\log_{10}$ (*P*-value). The green line indicates Bonferroni genome-wide significance, and the blue line is FDR significance.

- **Pretreatment**

There are three main types of pretreatments for revealing the presence or absence of the methyl group at cytosine residues (**Figure 9**). The first one is enzyme digestion, which relies on the fact that some restriction enzymes are inhibited by 5mC and 5hmC in the CpG context. Therefore, evaluating the patterns of cutting by such enzymes can provide

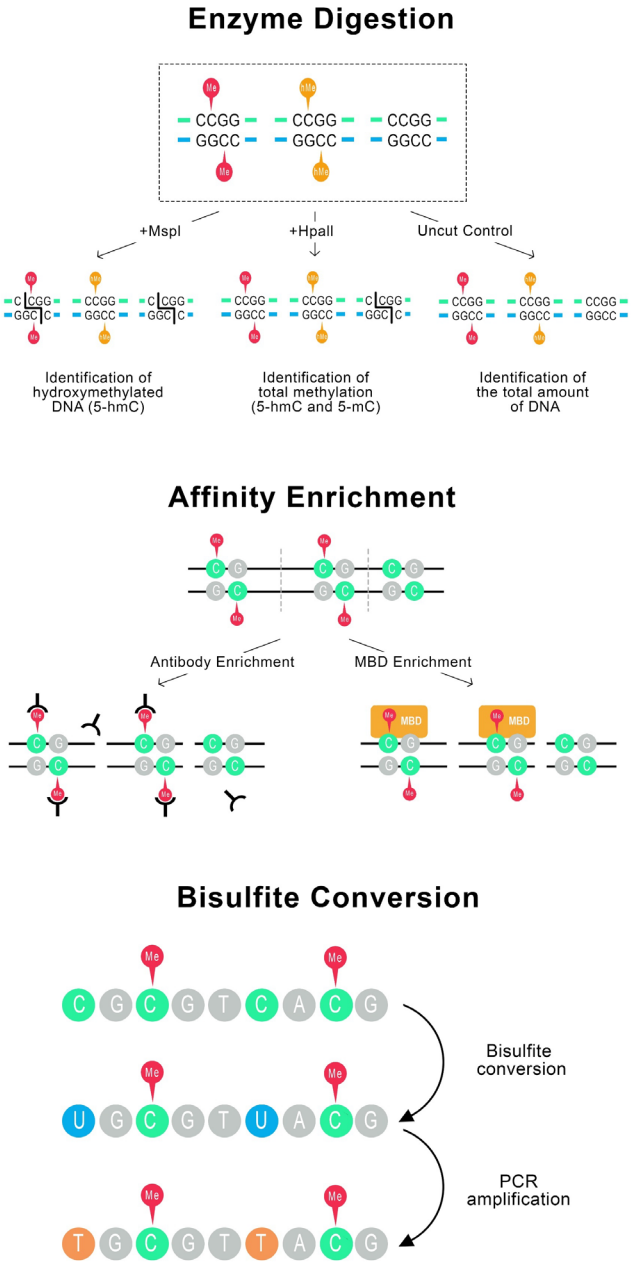


Figure 9 | The three main types of pre-treatments. At the top, enzyme digestion. In the middle, affinity enrichment. At the bottom, bisulfite conversion. Based on New England Biolabs, Lee et al., and Epigentek [79–81].

information about DNA methylation. This technique is very sensitive but extremely prone to false-positive results caused by incomplete digestion [75,76].

Second, affinity enrichment strategies use antibodies specific for 5mC (methyl-DNA immunoprecipitation - MeDIP) or methyl-binding proteins with affinity for methylated DNA. They are low-cost and straightforward for experienced laboratories but have relatively low resolution [75,77].

Third, bisulfite conversion is considered the gold standard pretreatment because of its single-nucleotide resolution, flexibility across organisms, and very low input requirements. The treatment with sodium bisulfite transforms unmethylated cytosines into uracils, whereas methylated cytosines are thermodynamically protected [78]. Then, bisulfite-converted fragments are subjected to PCR amplification, where uracils are replaced by thymines. The last step is to sequence the DNA amplified and compare it to a reference unconverted sequence to discover the sites that were originally methylated or not [77].

- **Analytical step**

Although the first analyses were performed with gel electrophoresis, nowadays microarrays and next-generation sequencing (NGS) techniques are broadly used when assessing DNA methylation. Microarrays are composed of bead-ligated probes that distinguish methylated and unmethylated loci based on their differential sequence [77]. The most extensively used arrays are based on Illumina probe extension: Infinium HumanMethylation450 (simplified as 450K) and

HumanMethylationEPIC (EPIC) BeadChips. While 450K was first developed and evaluates around 450,000 CpG sites epigenome-wide, the EPIC array assesses around 850,000 CpG sites. Besides, NGS techniques can be whole-genome bisulfite sequencing (WGBS) or reduced representation bisulfite sequencing (RRBS) when it is not necessary to measure the methylation status of every CpG. Sequence-based analyses have the advantage over array-based analyses in that they apply to any species for which a reference genome exists. A specific array of the specie of interest may be available to use the array-based methodology [75].

1.3. DNA methylation and genetic background

There is a high inter-individual variation in DNA methylation. One study revealed that approximately 50% of the CpG sites had more than 50% of variability across samples [82]. This is explained by differences in gender, age, environmental exposures, lifestyle, and ethnicity, but also due to genetic factors.

Genetic mutations are defined as any alteration in the DNA sequence. They are responsible for human evolution since they create variability over generations. When a sporadic mutation occurs in one individual, it can be inherited from their children. If the mutation has a positive effect, such as increasing the chance of survival or reducing disease risk, it may be inherited through generations and spread through the population. When this mutation occurs in at least 1% of the population, it is no longer considered a genetic mutation but a genetic variant.

Mutations can be inherited from the progenitors or can occur in the individual throughout life due to environmental exposures or errors during replication. Depending on the cells affected, we can find germline or somatic mutations. Germline mutations are those that occur in germline cells (cells that give rise to gametes), allowing to pass the mutation to the offspring. In this case, all the cells of the developing embryo will carry this mutation. Oppositely, somatic mutations occur in cells found elsewhere in an organism's body. These mutations are inherited by daughter cells through mitosis but not by the offspring (**Figure 10**).

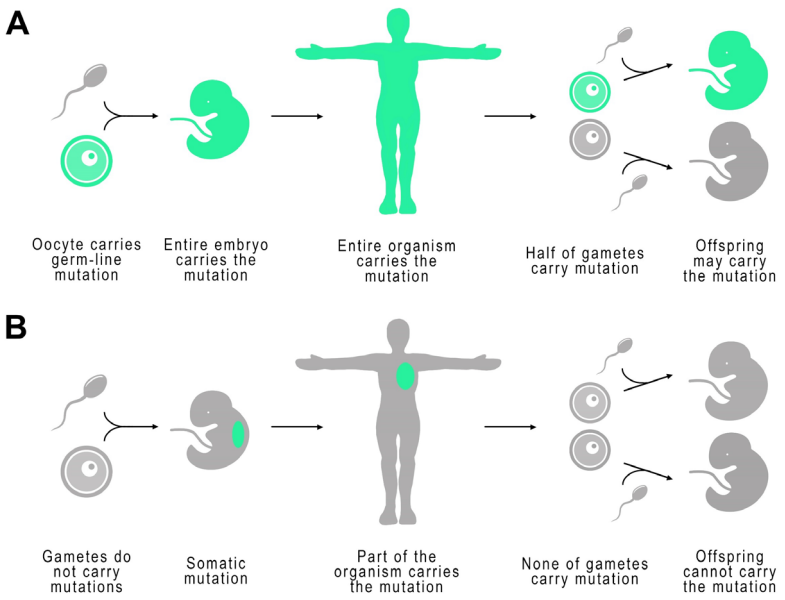


Figure 10 | Differences between germ-line and somatic mutations. (A) Germ-line mutation. (B) Somatic mutation.

Changes in the DNA sequence can occur at different levels. First, the most common variants are point mutations, which affect a single nucleotide. Second, mutations can also occur at the level of the chromosome, where large segments of the chromosome are affected. Third, genomic mutations are those affecting a whole chromosome.

1.3.1. Genetic mutations

Genetic mutations are the most common variants in humans and are defined as single nucleotide changes (**Figure 11**). These changes include the addition of a nucleotide (insertion), the removal of a

nucleotide (deletion), or the exchange of a nucleotide with another (substitution).

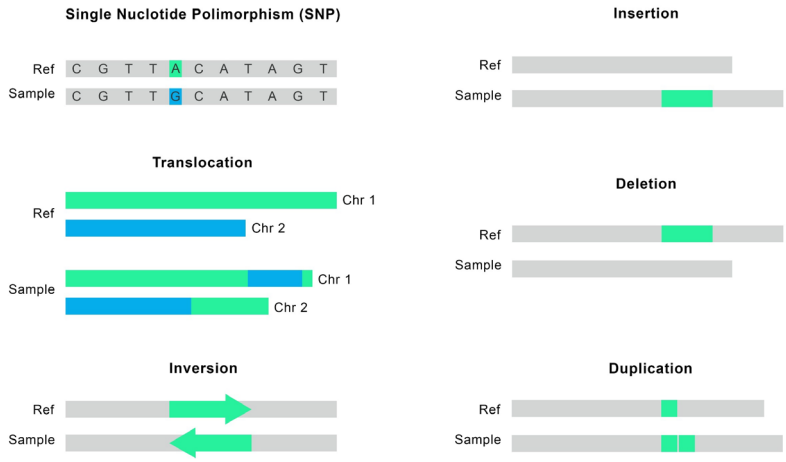


Figure 11 | Main types of genetic variants. While translocations and inversions are balanced structural variants, insertions, deletions, and duplications are unbalanced structural variants.

If genetic mutations occur in at least 1% of the population, they are considered single nucleotide polymorphisms (SNPs). Due to their prevalence, they are normal variants that do not directly cause disease. Although SNPs do not result in a disease, they can predispose individuals to certain traits or medical conditions. Therefore, SNPs are widely used as biomarkers to estimate the risk of disease. Typically, genome-wide association studies (GWAS) identify the most significant SNPs that are statistically associated with the phenotype of interest. This method consists of analyzing the genomes of many people and identifying the genetic variants that appear more frequently in those individuals with the trait of interest compared with those without the

disease or trait. Lastly, these variants are used to search for nearby variants that contribute directly to the disease or trait.

SNPs are one of the genetic factors most associated with DNA methylation changes. Many studies have identified a high percentage of CpG sites with *cis* and *trans* methylation quantitative trait loci (mQTL) in blood, lung, breast, kidney, and brain, among other tissues [83–86]. Moreover, the proportion of *trans*-mQTLs (more than 1 Mb between the SNP and the CpG site) is much smaller than the proportion of *cis*-mQTLs. Shi et al. remarked that mQTLs were enriched for DNase hypersensitive sites, modified histones, and binding sites. SNPs may affect these epigenetic marks by affecting the core recognition sequences, losing or gaining a CpG within a binding region (which methylation alters binding), or altering the binding sequence for interacting factors [86].

1.3.2. Chromosomal mutations

Chromosomal mutations are the second type of mutation and are also known as structural variants (SVs). They are defined as genetic alterations that affect a large region of a chromosome. When the alteration does not come with a loss of DNA, they are called balanced SVs. Among them, we can find translocations and inversions. Sometimes, chromosomal mutations lead to a loss of genetic material. In this case, SVs are considered unbalanced and include insertions, deletions, and duplications (**Figure 11**).

In contrast to SNPs, there are a few studies evaluating the effect of SVs on DNA methylation. However, Shanta et al. reported the influence on

3D DNA structure by large SVs, which can be highly correlated with changes at the DNA methylation level.

- **Translocations**

Translocations occur when chromosome breaks and the fragmented pieces re-attach to different chromosomes. They are classified as reciprocal and Robertsonian. Reciprocal translocations occur when two detached fragments from two non-homologous chromosomes are switched. This type of translocation occurs in about 1 in 491 live births [87]. Although they are usually harmless because the genetic material is balanced, their gametes may create unbalanced chromosome translocations during meiotic chromosomal segregation. On the other hand, Robertsonian translocations are caused by breaks at or near the centromeres of two acrocentric chromosomes. The reciprocal exchange leads to one large metacentric chromosome and one extremely small chromosome that may be lost during chromosome segregation. The most common Robertsonian translocation is between chromosomes 13 and 14 and has a prevalence of 0.97 every 1000 newborns [88].

There are a few studies investigating the effect of translocations on DNA methylation. McCartney et al. evaluated the effect of the balanced translocation between chromosomes 1 and 11, which is linked to major mental illness, on DNA methylation [89]. They found a differential DNA methylation at the regions surrounding the translocation breakpoints, that was, indeed, implicated in neuronal development and psychiatric illness.

- **Inversions**

Inversions are segments of DNA that run in the opposite direction to a reference genome. An inversion occurs when a single chromosome undergoes breakage and rearrangement within itself. Inversions can be classified as paracentric and pericentric depending on whether they involve or not the centromere. Pericentric inversions include the centromere, and the breakpoints are one in each arm. Paracentric inversions have both breakpoints in the same arm, excluding the centromere from the inversion.

Many inversions are identified in a huge proportion of the population since they do not directly cause disease. The three most common inversions in humans are large and are located at 8p23.1, 16p11.2, and 17q21.31 cytogenetic regions (**Table 2**). Like SNPs, inversions are important contributors to the genetic basis of common complex diseases in humans. Recent studies, most of them carried out at the BRGE lab, demonstrated that inversions are associated with an increased risk of obesity, diabetes, asthma, cancer, and neurological conditions [90–97].

Although several studies have demonstrated the effect of inversions on gene expression, it is unknown the extent to which inversions are also characterized by specific methylation patterns. However, some studies have already reported associations between inversion and phenotypes likely modulated by specific methylation changes [92,98,99]. For instance, Ruiz-Arenas et al. reported that the effect of the inversion 17q21.31 on colorectal disease-free survival is more likely mediated by DNA methylation that by gene expression [92].

Table 2 | Characteristics of the three most common polymorphic inversions in humans (8p23.1, 16p11.2, and 17q21.31). The table shows the length in kb, the mapping coordinates (hg19), and the frequency of all inversions obtained from *scoreInvHap* [97].

Genomic inversion	Length (kb)	Inversion region	Inversion frequency (%)
8p23.1	3,924.86	chr8:8,055,789-11,980,649	57.95
16p11.2	364.17	chr16:28,424,774-28,788,943	34.49
17q21.31	710.89	chr17:43,661,775-44,372,665	23.96

- **Copy number variants**

Copy number variants (CNVs) occur when the number of copies of a specific segment of DNA varies among different individuals' genomes. The size of a CNV is typically larger than 50 bp, whereas smaller elements are known as insertions and deletions (indels) [100]. CNVs may account for almost 10% of an individual's genome [101]. On some occasions, the segment with a different number of copies can include one or more genes. When this gene encodes for a protein, the variation in the number of copies can result in a variation in the amount of the specific protein. Furthermore, CNVS can also regulate gene expression by altering regulatory elements associated with gene expression.

A recent study by Shi et al. demonstrated that germline inter-individual CNVs are associated with differences in DNA methylation at the CpG level. Indeed, these associated mQTL-CpG patterns are correlated with transcript expression, are enriched for regulatory elements, and are located in previously reported disease risk loci [102]. Another study

demonstrated that somatic CNVs in cancer are associated with DNA methylation levels [103]. However, the causality is unclear since there are other studies suggesting that methylation changes may lead to CNV formation due to the increase in DNA breakage [104].

1.3.3. Genomic mutations

The normal karyotype is organized into 46 chromosomes: 22 pairs of autosomal chromosomes and two sexual chromosomes (two X chromosomes in females and an X and a Y chromosome in males). Genomic mutations, also known as aneuploidies, involve the alteration of the number of chromosomes, caused by the addition or missing of one or more chromosomes. The most common aneuploidy in humans is the presence of an additional chromosome (trisomy). Although they represent 0.3% of all live births, they also account for 35% of spontaneous abortions since most of them are not compatible with life [105].

The most common viable trisomy involves chromosome 21 and it is known as Down Syndrome. Muskens et al. found important genome-wide effects on DNA methylation in hematopoietic cells in newborns with Down syndrome. Interestingly, the most significant CpG sites mapped to two genes with important roles in the regulation of hematopoietic development. They suggested a contribution of DNA methylation in the high prevalence of hematological problems in children with Down syndrome [106].

1.4. DNA methylation and environment

As mentioned previously, the epigenetic landscape of genes can be altered due to environmental exposures. In some cases, these modifications may lead to disease [37–39]. Many studies demonstrate that the environment and genetic predisposition are important contributors to the risk of chronic diseases. In many cases, it is even more important the role of the environment compared with the role of genetic causes. In 2005, Wild contrasted the efforts and the costs invested in genetic studies and environmental studies. Clearly, genetic studies were more developed, with sophisticated techniques, in contrast to environmental studies which did not evolve too much since the 70s.

1.4.1. The exposome

To avoid such inequality, Wild introduced the term "exposome", which did not include only hazard exposures, but all the exposures (internals and externals) to which an individual was subjected, from conception onwards [107]. This term includes information about the personal exposome, the external exposome, and the biological responses of an individual (**Figure 12**).

The personal exposome does not only encompasses exposures to toxic chemicals, drug intake, smoking, and alcohol, but also information about the diet, physical activity, hours of sleep, type of work, and social life.

Figure 12 | The exposome. This figure shows the different types of exposomes and examples for each of them. In green, the personal exposome. In blue, the external exposome. In orange, the biological responses. Image from ISGlobal (https://www.isglobal.org/exposome-hub).

The external exposome is the most similar to the traditional concept of environment. This includes exposure to air pollution, noise, traffic, light, sunlight, and greenspaces, among others. In this group, it is also encompassed whether the individual leaves in an urban environment, defined as an area with an increased density of human-created structures in comparison to the areas surrounding it.

Biological responses are maybe the most forgotten part of the exposome. However, they are very important since they have an important role in disease emergence. Among these internal exposures, it is remarkable the importance of the gut microbiome, the stress, the inflammation, the

aging, and the different omics involved in the biological pathways (gene expression, epigenetics, proteomics, and metabolomics).

1.4.2. Modulation of DNA methylation by the exposome

As mentioned previously, the exposome consists of a group of internal and external exposures. In this section, the effect of the most important exposures on DNA methylation will be described.

- **Toxic chemicals**

Many chemicals are known to be carcinogenic. For a long time, it was considered that the carcinogenic effect was exclusively due to the alteration of DNA. In 1991, Ashby and Tennant found that 162 out of 301 chemicals tested were carcinogens, and only 98 of them were genotoxic [108]. This illustrates that carcinogenesis may occur in the absence of mutagenesis, through epigenetic modifications [39].

Bisphenol-A (BPA) is an endocrine disruptor and is one of the most used chemicals worldwide. It can be found in bottles, plastic containers, and cans in which drinks and food are stored. Among the consequences of BPA exposure, it is remarkable the gene silencing by CpG promoter methylation in breast cancer epithelial cells [109], and the increased susceptibility to preneoplastic prostate changes during aging by epigenetic mechanisms [110].

Most carcinogenic metals are characterized by low mutagenic activity and potential epigenetic mechanisms leading to cancer [39]. As an example, nickel compounds are not known as mutagens but have carcinogenic effects. In this case, Ni^{2+} can substitute Mg^{2+} in the DNA

phosphate backbone and increase chromosome condensation. This establishes heterochromatin regions where gene expression is silenced. When these regions contain suppressor genes, their inactivation may lead to cancer [111]. Cadmium is another metal highly associated with cancer and with low levels of mutagenicity. High concentrations of cadmium are correlated with DNMTs inhibition, leading to global hypomethylation [112]. Studies in arsenic have revealed global DNA alterations and gene promoters' methylation levels, histone acetylation and phosphorylation, and miRNA expressions. Those effects have been linked to carcinogenesis [113].

- **Lifestyle**

It is already known that physical activity is an important contributor to a better quality of life. In part, this can be attributed to the epigenetic changes that are produced after physical exercise. For example, acute exercise is associated with hypomethylation of many promoters of metabolic genes, resulting in higher expression [114]. Besides, chronic moderate exercise protects the cells from the inflammatory environment by preventing the activation of inflammatory cytokines [115]. Finally, physical activity has been linked to DNA methylation patterns that result in oncogene silencing and higher expression of tumor suppressors [116,117].

Some recent studies have shown that people with post-traumatic stress disorder present different levels of DNA methylation compared with controls [118]. More interestingly, children from mothers who had stress during gestation were more likely to have psychiatric disorders due to the high expression of glucocorticoid receptors [119].

The sleep pattern also influences methylome. Lahtinen et al. performed an EWAS comparing individuals who self-reported insufficient sleep versus controls. They found that 78% of the differentially methylated positions (DMP) were hypomethylated in cases. This suggests that there is a global hypomethylated in individuals with sleep disorders. Additionally, they found 12 DMPs in a region that was previously associated with Smith-Magenis syndrome, which consists of a rare condition that comprises disturbed sleep and inverse circadian rhythm [120].

Tobacco, alcohol, and marijuana are highly consumed worldwide. Due to their association with the risk of mortality and many health conditions, the high consumption of these drugs has become a public health problem. Recent studies suggest that epigenetics is a potential mediator between drug consumption and disease risk. Smoking is one of the most associated exposures with DNA methylation. Several studies have demonstrated the high number of CpG sites differentially methylated along the genome [121–124], even when the exposure was during pregnancy [125]. On the other hand, alcohol is also associated with genome-wide DNA methylation. Compared with tobacco, there are more CpG sites differentially methylated but with lower effects [126–130]. Oppositely to tobacco and alcohol, a few studies have studied the effect of marijuana on DNA methylation and they have revealed small effects [131,132].

- **Nutrition**

The impact of nutrition during development has been widely studied. During World War II, there were a lot of people suffering from hunger,

especially pregnant women. Heijmans et al. demonstrated that individuals who suffered prenatal hunger showed low DNA methylation levels six decades later in *IGF2*, a gene implicated in growth development. Also, these individuals were more likely to develop a subset of diseases, including schizophrenia, stress sensitivity, and obesity [133].

Many studies have evaluated the effect of dietary factors on DNA methylation in blood samples. One of the main studied factors is folate intake, which has revealed controversial associations depending on the study. While Stidley et al. and Ono et al. found a negative association between folate intake and global DNA methylation, Zhang et al. and Agodi et al. found a positive association between folate intake and LINE-1 methylation [134–137]. Moreover, Barchitta and colleagues brought to light a positive association between the Mediterranean diet and the LINE-1 methylation level [138]. Interestingly, the level increased in individuals more adhered to healthy dietary patterns. Another study demonstrated that the consumption of fruit was associated with differentially methylated CpG sites that mapped to genes involved in antigen presentation and chromosome and telomere maintenance [139].

- **Air pollution**

Recent studies suggest that air pollutants alter epigenetic mechanisms. Particulate matters (PM) consist of solid and liquid particles suspended in the air. According to their size, they can be classified as coarse (diameter 10μm; PM10), fine (diameter 2.5μm; PM2.5), and ultrafine (0.1μm; PM0.1) particles [140]. It is demonstrated that PM2.5 and

PM10 exposures are associated with the hypomethylation of Alu and LINE-1 repetitive elements in leukocytes and buccal cells [141]. Additionally, PM2.5 exposure during pregnancy is associated with global levels of methylation in the placenta [142]. Besides, high exposure to PM2.5 has been seen associated with hypomethylation of many tumor suppressor promoters (such as p53, p15, and p16), which can be explained by the overexpression of DNMT1 [143]. Other epidemiological studies reveal that PM exposure modulates the methylation of several inducible nitric oxide synthase (iNOS) genes, which are important for asthma pathogenesis [144]. In line with these findings, exposures to NO_2 and CO alter DNA methylation at genes implicated in asthma development [145].

1.4.3. Polyenvironmental risk scores

Typically, epidemiological studies evaluate individual environmental risk factors to establish the risk of a phenotype of interest. An environmental risk factor is defined as any exposure from the personal exposome, external exposome, or biological responses that could predispose an individual to disease. Sometimes, individual exposures are not enough to predict a disease by itself. A few studies have suggested that an aggregate score representing multiple environmental risk factors may predict the development of a medical condition.

In 2018, Padmanabhan et al. introduced the term “polyenviromic risk score” (PERS), analogous to the polygenic risk score used in genetics [146]. In their research, they used PERS to predict the conversion to psychosis since individual environmental risk factors were insufficient to predict the risk. Early identification of individuals at risk of psychosis

is very important since the disease may be delayed or even prevent the onset with interventions before the full manifestation [147]. Therefore, their PERS included nine risk factors that were known to increase the risk of psychosis independently in previous studies.

Following the basis of the polygenic risk score (PRS), a list of odds ratios for each risk factor's association with psychosis was obtained from the literature. Similar to the number of risk alleles for each SNP in PRS (either 0, 1, or 2), each risk factor was binarized. Then, for an individual, the log of the odds ratio for each exposure was multiplied by either 1 (risk factor is present in the individual) or by 0 (absent). These products were added together, and the sum was divided by the total of environmental risk factors assessed (nine in their case). As expected, the PERS score was significantly correlated with conversion to psychosis.

More recent studies have applied similar methodologies to estimate aggregate scores that represent multiple environmental risk factors. However, most of them are focused on psychosis. For instance, Jeon et al. developed a Korea-PERS (K-PERS) for psychosis [147]. It is worth mentioning that, in this paper, PERS stands for "polyenvironmental risk score" instead of "polyenviromic risk score". For creating the K-PERS, they followed the methodology described in Oliver et al. and Vassos et al. [148,149]. Instead of a weighted sum of odds ratio, they used the relative risks (RR) associated with each factor. They first estimated a raw score for each factor as the 10-base logarithm of its RR. For instance, the raw score of the urbanity factor is $\log_{10}(2.2) = 0.34$ because the estimated RR of psychosis in individuals in the urban setting is 2.2. Second, they subtracted the population average of this raw score, resulting in positive scores for those individuals at risk and negative

scores for those not at risk. Third, they multiplied the subtracted scores by 10 and rounded them to the nearest half-integer for easy use. Using this methodology, the K-PERS was developed, and it was able to distinguish between patients with schizophrenia spectrum disorders and healthy controls.

Considering that these tools are achieving positive results, similar methods should be applied to other diseases in which early detection could improve the prevention and monitoring of the medical condition.

1.5. Gene-environment interaction

1.5.1. Definition

So far, we have described the effects of genetic factors and environmental exposures on DNA methylation independently. However, there are many situations where the effect of an environmental factor on a phenotype varies in different populations, as well as the effect of a genetic variant may differ depending on the environment. This is what is known as gene-environment (GxE) interaction and it refers to the modification of the effect of a genetic variant on a phenotypic trait by an environmental factor, and vice versa, the modulation of the effect of environmental exposure on a phenotypic trait by genetic factors [150].

The first examples of GxE interactions were described by Caspi et al. They showed that antisocial behavior differed according to the interaction between childhood experiences of maltreatment (none, probable, or severe) and a genetic variant in the *MAOA* gene, characterized by synthesizing high or low levels of the monoamine oxidase A (MAOA) enzyme [151]. Similarly, Caspi et al. demonstrated that the variation in depression was associated with the interaction between the number of stressful life events experienced (0, 1, 2, 3, or 4 +) and a polymorphism in the gene *5-HTT* involved in the regulation of serotonin transporters production (high, moderate, or low levels) [152]. In both cases, the phenotype (antisocial behavior or depression) was best predicted by considering the interaction between the genetic factor and the environmental factor than considering the genetic factor alone or the environmental factor alone (**Figure 13**).

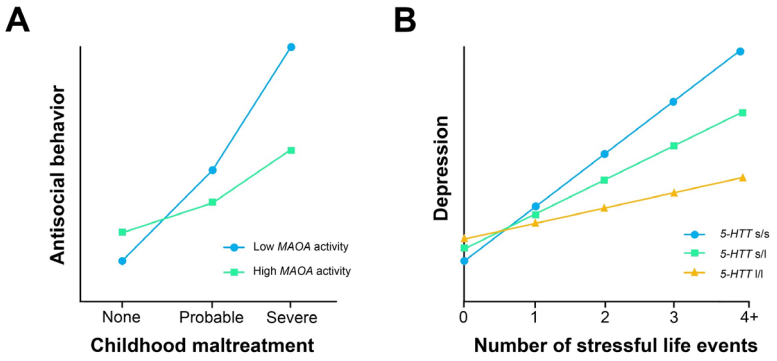


Figure 13 | Two examples of significant gene-environment interactions. (**A**) Regression analysis estimating the association between the childhood history of maltreatment and antisocial behavior as a function of *MAOA* activity. Figure adapted from Caspi et al. [151]. (**B**) Regression analysis estimating the association between the number of stressful life events and depression as a function of the *5-HTT* genotype. Figure adapted from Caspi et al. [152].

GxE interactions can be useful to modulate the adverse effects of a risk allele by avoiding exposure to a specific hazard. Therefore, GxE interactions result in a modifiable relationship between genetics and phenotype-associated changes [150]. Those interactions can reduce the effect of a genetic variant if the environmental risk factor is limited in a study population and if the genetic variant has small effects outside that environment. Otherwise, individuals living in environments enriched in this exposure will show earlier the effects of the genetic variant. Therefore, genetic influence may differ depending on the environment [153].

Uher realized that there was another reason that justified GxE interactions [154]. Individuals with common psychopathologies, like major depression or anxiety, usually have a small reproductive

disadvantage. Along with this, we would expect a negative selection to remove a harmful genetic variation. When GxE interactions occur, many individuals are carrying the risk allele without developing the disease because they are not exposed to the environmental risk factor. Additionally, the same allele may confer a reproductive advantage in other environments or circumstances. Then, the risk allele could persist in the population endlessly.

1.5.2. DNA methylation role in GxE interactions

Environmental and genetic factors are thought to modulate the epigenome to determine the phenotype or disease risk [155–157]. As mentioned previously, it is explained by the direct effect of environmental exposures on epigenetics, mainly on DNA methylation, which is associated with changes in gene expression that can explain, lastly, the effect on the trait of interest. Moreover, genetic factors also have an important role in regulating the epigenome. Therefore, DNA methylation is the best intermediary in the relationship between GxE interactions and diseases.

Czamara et al. evaluated the relative contributions of environmental and genetic factors on DNA methylation in neonatal blood in variably methylated regions [158]. They tested which factors best explained the variability of methylation: genotypes (G), environmental factors (E), the additive effect (G+E), or the interaction effect (GxE). Their results showed that GxE models were the best predictors of DNA methylation variance. Interestingly, the enrichment of the SNPs implicated in the significant GxE interactions was enriched in common disorders. This

suggests that genetic variants alter the effect of environmental exposures at the DNA methylation level and the disease risk.

1.5.3. Limitations in GxE interaction research

One of the main limitations of GxE research is the limited replication and vulnerability of publication bias. The statistical power needed to detect GxE interactions depends on many factors. First, the detection of GxE interactions requires 4 times the sample size needed to find genetic main effects. Second, the distribution of the individuals in the different environments affects the statistical power. If a genetic factor has the main effect in environment A and not in environment B, the interaction will be found more easily if the samples are equally distributed in both environments. Oppositely, if all the samples are from one of the environments, the interaction cannot be detected. Third, the power to detect interactions increases when either the quality of measurement of the phenotype or the environmental risk factor is high. Fourth, it is easier to detect interactions when the gene function is more proximal to the phenotype or biological process.

To sum up, low statistical power can result in a failure to replicate or detect an interaction. Nowadays, the most accepted methodology for finding interactions is the use of meta-analysis, since it requires multiple studies that test the same hypothesis.

2 | HYPOTHESES AND OBJECTIVES

2.1. Hypotheses

DNA methylation is a dynamic process that has an important role in the regulation of gene expression. Since abnormal DNA methylation can result in common diseases, evaluating the genetic and environmental factors that alter DNA methylation by themselves or by gene-environment interactions is crucial for developing new biomarkers.

2.1.1. Hypothesis 1

- **General hypothesis**

DNA methylation is altered by the exposome.

- **Specific hypotheses**

❖ Lifestyle factors, such as tobacco, alcohol, and marijuana consumption, are potential modulators of DNA methylation along the genome.

❖ The association of drug consumption with DNA methylation may partially explain the association between drug use and common diseases.

2.1.2. Hypothesis 2

- **General hypothesis**

Genetic factors and gene-environment interactions are strong modulators of DNA methylation.

- **Specific hypotheses**

❖ Polymorphic inversions, a type of structural variant, may be associated with DNA methylation patterns because of their large size that englobes multiple genes and the association of inversions with gene expression and common diseases.

❖ The effect of environmental factors on DNA methylation may be different according to the inversion genotype.

2.1.3. Hypothesis 3

- **General hypothesis**

The aggregation of multiple environmental factors can be used to predict sexual dimorphism in disease risk and DNA methylation can participate as a mediator.

- **Specific hypotheses**

❖ The combination of multiple exposures during pregnancy may identify an environment with different risk of obesity between boys and girls.

❖ Differences in obesity risk between boys and girls can be associated with differences in neurodevelopment.

❖ DNA methylation can be used to determine whether one individual belongs to the environment with high sexual dimorphism in obesity.

2.2. Objectives

This thesis aims to find novel modulators of DNA methylation that may be used later to develop biomarkers for disease development using large populations.

2.2.1. Objective 1

- **General objective**

To detect differentially methylated CpG sites along the genome associated with drug consumption to explain the association with common diseases.

- **Specific objective**

❖ To identify blood DNA methylation changes associated with the level of consumption of tobacco, alcohol, and marijuana by carrying out three independent epigenome-wide association studies. **Manuscript 1**

❖ To identify new DNA methylation loci that mediate the effect of drug consumption on cardiovascular disease risk, especially in hypertension. **Manuscript 1**

2.2.2. Objective 2

- **General objective**

To assess the effect of large structural variants on DNA methylation and the interaction with the exposome.

- **Specific objective**

❖ To evaluate whether three common polymorphic inversions in humans (8p23.1, 16p11.2, and 17q21.31) affect the methylation patterns of their encapsulated and surrounding DNA sequences in blood cells from children. **Manuscript 2**

❖ To test whether the same DNA methylation patterns are detected in prenatal heart tissue. **Manuscript 2**

❖ To assess whether a large set of 64 early life exposures had a different impact on DNA methylation according to the genotype of 8p23.1, 16p11.2, and 17q21.31 inversions. **Manuscript 2**

2.2.3. Objective 3

- **General objective**

To identify an environment with high sexual dimorphism in obesity and neurodevelopment detectable by DNA methylation changes.

- **Specific objectives**

❖ To recognize an environment consisting of a multiexposure profile during pregnancy where girls are more protected than boys against childhood obesity and neurodevelopment delay. **Manuscript 3**

❖ To identify a DNA methylation pattern associated with the specific environment that affects different the risk of obesity in boys and girls. **Manuscript 3**

❖ To infer subpopulations with high sexual dimorphism in obesity and academic achievement in an independent cohort. **Manuscript 3**

3 | STUDY POPULATIONS

To address the objectives of this thesis, we used data from three different study populations. In the following lines, I will briefly describe these projects.

3.1.1. HELIX project

The Human Early Life Exposome (HELIX) project [159] comprises a total of 1,301 mother-child pairs from six birth cohorts in Europe: BIB (Born in Bradford; the United Kingdom) [160], EDEN (Etude des Déterminants pré et postnatals du développement et de la santé de l'Enfant; France) [161], INMA-SAB (Infancia y Medio Ambiente; Spain; subcohort Sabadell) [162], KANC (Kaunas cohort; Lithuania) [163], MoBa (The Norwegian Mother, Father and Child Cohort study; Norway) [164], and Rhea (Greece) [165].

These mother-child pairs participated in a common, completely harmonized, follow-up examination between December 2013 and February 2016, when children were between 6-11 years old.

The main goal of this project was to implement exposure assessment and biomarker methods to characterize early-life exposure to multiple environmental factors [166] and associate these with omics biomarkers and child health outcomes.

For these same children, multi-omics molecular phenotyping was performed, including measurement of blood DNA methylation (450K, Illumina), blood gene expression (HTA v2.0, Affymetrix), blood miRNA expression (SurePrint Human miRNA rel 21, Agilent), plasma proteins (Luminex), serum metabolites (AbsoluteIDQ p180 kit,

Biocrates), urinary metabolites (1H NMR spectroscopy), and DNA microarray (Chemagen kit, Perkin Elmer).

In the INMA cohort, the children were re-contacted between 2021 and 2022 when they were around 15 years old. During my PhD, I participated in the blood and urine samples processing in AIRLab laboratory within ISGlobal (www.isglobal.org).

3.1.2. TruDiagnostic DNA biobank

The TruDiagnostic DNA biobank includes 3,890 individuals recruited between October 2020 and February 2022. Those individuals have taken the commercial TruDiagnostic TruAge test [167] and methylation data has been generated from them (EPIC, Illumina). As this testing is priced to the consumer at approximately $500, this study cohort is relatively more affluent than random sampling or traditionally banked cohorts. Additionally, most of these individuals tend to be seeking preventive medicine and have fewer comorbidities than normal patient populations, which is known as the healthy donor effect.

This biobank includes individuals from the EEUU aged between 13 and 97 years old. Information about their demography, lifestyle, and history is reflected in the questionnaire each participant filled in.

3.1.3. Biobanc Hospital Universitari Vall d'Hebron

To test whether the DNA methylation patterns in polymorphic inversions were detected in prenatal heart tissue, we used samples from the Hospital Universitari Vall d'Hebron (HUVH) Biobank. We had

access to human fetal samples from 40 fetuses of terminated pregnancies due to a major congenital heart defect (gestational age 21-22 weeks in all cases). Heart tissue DNA was obtained following necropsy using standard procedures, whole genome sequencing was performed at Centogene, and DNA methylation was measured with Infinium MethylationEPIC [168].

4 | KNOWLEGDE TRANSFER

Collaboration with the TruDiagnostic company

The PhD has provided me with the opportunity to collaborate with a US company, TruDiagnostic, thanks to a formal collaboration the company has with my PhD supervisor. TruDiagnostic is a start-up company that uses a multi-omics approach to help scientists, physicians, and patients to understand and benefit from the insights found in the fluid epigenome.

Mainly, they offer a test to calculate the biological aging of their patients based on epigenetic aging algorithms. They provide a complete report that is personalized for each patient and summarize the suggestions to improve their biological aging based on validated research.

At the same time, the company is in constant growth and collaborates with private and public institutions to improve their knowledge of epigenetics and give better suggestions to their patients. We had the opportunity to collaborate with them in several lines, some of them related to my PhD. First, we created a pipeline for the pre-processing of DNA methylation data, as well as for phenotype data. Then, we evaluated the impact of multiple drugs on DNA methylation and the mediation between drug consumption and hypertension, which resulted in very interesting findings. In another study, we used TruDiagnostic participants for testing the predictor of a prenatal environment based on methylation.

Collaborating with TruDiagnostic allowed me to know the way private companies operate, particularly in the case of start-ups. In the case of TruDiagnostic, major efforts are made to transfer scientific research to the general public. In this line, I have participated in the creation of an audiovisual material to explain why the study of DNA methylation is important to people's health and how it can be used to prevent diseases.

5 | MANUSCRIPTS

5.1. Manuscript 1

Impact of tobacco, alcohol, and marijuana on genome-wide DNA methylation and its relationship with hypertension

Carreras-Gallo N, Dwaraka VB, Cáceres A, Smith R, Mendez T, Went, H, González JR

Impact of tobacco, alcohol, and marijuana on genome-wide DNA methylation and its relationship with hypertension. *Epigenetics.* Under review. IF: 4.9. Position: Q1

Supplementary Material here

ARTICLE

Impact of tobacco, alcohol, and marijuana on genome-wide DNA methylation and its relationship with hypertension

Natàlia Carreras-Gallo[1], Varun B. Dwaraka[2], Alejandro Cáceres[1,3,4], Ryan Smith[2], Tavis L. Mendez[2], Hannah Went[2], Juan R González[1,3,5 *]

Abstract

Background: Tobacco, alcohol, and marijuana consumption is an important public health problem because of their high use worldwide and association with the risk of mortality and many health conditions, such as hypertension. High blood pressure is the commonest risk factor for death throughout the world. A likely pathway of action of substance consumption leading to persistent high blood pressure is DNA methylation. Here, we evaluated the effects of tobacco, alcohol, and marijuana on DNA methylation in the same cohort (N=3,424). Three epigenome-wide association studies (EWAS) were assessed in whole blood using the InfiniumHumanMethylationEPIC BeadChip. We also evaluated the mediation of the top CpG sites in the association between substance consumption and hypertension.

Results: Our analyses showed 2,569 CpG sites differentially methylated by alcohol drinking and 528 by tobacco smoking. We did not find significant associations with marijuana consumption at Bonferroni level. We found 20 genes overlapping between alcohol and tobacco that were enriched in signaling functions of the nervous system and neurodevelopment. In the mediation analysis, we found 8 CpG sites that significantly mediated the effect of alcohol consumption on hypertension. The top alcohol-related CpG site (cg06690548, *P*-value = $5.9 \cdot 10^{-83}$) mapped to

[1] Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain. [2] TruDiagnostic, Lexington, KY, United States. [3] Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [4] Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya, Barcelona 08019, Spain. [5] Department of Mathematics, Universitat Autònoma de Barcelona, Bellaterra, Spain. * Email: juranr.gonzalez@isglobal.org

SLC7A11 and strongly mediated the 73.6% of the effect of alcohol consumption on hypertension (*P*-value = 0.008).

Conclusions: Our findings suggest that DNA methylation should be considered for new targets in hypertension prevention and management, particularly in relation to alcohol consumption. Our data also encourage further research into the use of methylation in blood to study the neurological effects of substance consumption.

Keywords: Tobacco, Alcohol, Marijuana, DNA methylation, Epigenome-wide association study, Hypertension

Introduction

Tobacco, alcohol, and marijuana are the most commonly used drugs of abuse in the United States (1). While tobacco and alcohol consumption is legal, marijuana is the most commonly used illicit drug globally (2). The consumption of these substances is increasing, mainly among adolescents, and the health and social problems associated with them are an important public health concern.

Cigarette smoking is the leading cause of preventable death and disease in the US and is responsible for approximately 8 million worldwide deaths every year (3). Most of these deaths arise from cancers (mainly lung cancer), respiratory disease, and cardiovascular disease (4). Light to moderate alcohol intake is associated with reduced risks for total mortality, cardiovascular disease, and diabetes. However, excessive alcohol is the third leading cause of premature death in the US (5). Heavy alcohol use is associated with a higher risk of cardiovascular disease, diabetes, cirrhosis of the liver, pancreatitis, and cancer (6). Among marijuana health impacts, disturbances in the level of consciousness, cognition, perception, affect or behavior, and other psychophysiological functions and responses are known as short-term effects. Additionally, long-term marijuana consumption can increase the risk of dependence, cognitive impairment, mental disorders (psychoses, depression, anxiety, and suicidal behavior), and adverse physical health effects such as cardiovascular disease, chronic obstructive pulmonary disease, and respiratory and other cancers (2).

Many of the pathways whereby tobacco, alcohol, and marijuana exert adverse effects on health outcomes are unclear. Recent research suggests that epigenetics is a potential mediator between the consumption of toxic substances and the increase in common disease risk (7–10). DNA methylation, the most studied epigenetic modulation, consists of the

addition of a methyl group (-CH3) in the cytosine nucleotide without changing the DNA sequence. It occurs in the context of CpG sites, which are defined as adjacent cytosine and guanine nucleotides by a phosphate group. DNA methylation is dynamic and can be modified by genetic factors, disease, environmental exposures, and lifestyle (10–12). Moreover, DNA methylation can change during the lifetime and across tissues and cell types (13,14).

Although genetic mechanisms have been the focus of understanding human diseases, the disruption of the epigenetic balance can result in the modulation of gene expression. Consequently, epigenetic disruption can cause several major pathologies, including cancer and cardiovascular disease (7). Tobacco smoking is one of the exposures with a higher impact on DNA methylation, even when the mother smoked during pregnancy (15–21). Many studies also demonstrate that alcohol consumption produces methylation changes at the CpG site level (22–27). Conversely, only a few studies have demonstrated the effects of marijuana consumption on DNA methylation, all of which have shown small effects (28,29).

Hypertension, also known as high blood pressure, is a medical condition in which the blood pressure in the arteries is persistently elevated. It affects one billion people and is the most common risk factor for death worldwide (30). There are many factors associated with a higher risk of hypertension, including body mass index, tobacco use, physical activity, and alcohol consumption, among others (31). Light to moderate alcohol consumption seems to protect against hypertension because it decreases systolic and diastolic blood pressure. However, excessive intake accounts for about 16% of cases of hypertension worldwide (32,33). Cigarette smoking enhances hypertension by inducing cardiovascular mitochondrial oxidative stress (34,35). On the other hand, some studies evaluating the effect of marijuana consumption on blood pressure have revealed different results. For

instance, Abuhasira et al. demonstrated the therapeutic effect of marijuana in reducing blood pressure in hypertensive patients (36). However, other studies revealed an increase in blood pressure after marijuana consumption (37,38). In the light of the above, we hypothesized that changes in DNA methylation produced by substance consumption may partially explain its relationship with hypertension.

To this end, in this study we aimed to: i) perform a genome-wide association study of DNA methylation with tobacco, alcohol, and marijuana consumption; ii) identify the physiological pathways whose methylation is affected by those drugs; iii) evaluate the mediation between substance consumption and hypertension by methylation changes at the CpG site level.

Methods

The Study Cohort

Our study sample included 3,590 individuals from the TruDiagnostic DNA biobank recruited between October 2020 and February 2022. Those individuals have taken the commercial TruDiagnostic TruAge test and methylation data was generated from them. This is an EEUU population-based cohort aged between 13 and 97 years old. Among them, 58.7% are male. Demographic and substance use characteristics of the samples that met the QC requirements (N = 3,424) are displayed in Table 1. As this testing is priced to the consumer at approximately $500, this study cohort is relatively more affluent than random sampling or traditional banked cohorts. Additionally, most of these individuals tend to be seeking preventative medicine and have fewer comorbidities than normal patient populations, which is known as the healthy donor effect.

Table 1 Characteristics of participants in the TruDiagnostic DNA Biobank.

	N=3424
Sex, male	2010 (58.7%)
Age in years, mean (range)	52.9 (13.3 – 97.8)
Ethnicity	
European	2584 (75.5%)
African American or Black	70 (2.0%)
Asian	41 (1.2%)
Latino or Hispanic	276 (8.1%)
Middle Eastern or North African	76 (2.2%)
Native American or Alaska Native	26 (0.8%)
Pacific Islander or Oceanian	23 (0.7%)
Sub-Saharan African	7 (0.2%)
Other	321 (9.4%)
BMI (kg/m^2), median (range)	25.4 (10.1-58.2)
Tobacco consumption	
None	3275 (95.6%)
Less than 1 cigarette per week	48 (1.4%)
Less than 1 cigarette per day	25 (0.7%)
1-5 cigarettes per day	27 (0.8%)
6-10 cigarettes per day	21 (0.6%)
11-20 cigarettes per day	20 (0.6%)
More than 20 cigarettes per day	8 (0.2%)
Alcohol consumption	
Never	634 (18.5%)
On special occasions	976 (28.5%)
Once per week	578 (16.9%)
3-5 times per week	794 (23.2%)
Regularly	442 (12.9%)
Marijuana consumption	
Missing	149
Never	2908 (88.8%)
On special occasions	180 (5.5%)
Once per week	46 (1.4%)
3-5 times per week	73 (2.2%)
Regularly	68 (2.1%)

All the continuous variables are shown as mean (range) and the categorical variables as n (%).
BMI: body mass index

DNA methylation Assessment

Peripheral whole blood was collected by the lancet and capillary method into lysis buffer and DNA extract, and 500 ng of DNA of bisulfite were converted using the EZ DNA Methylation kit (Zymo Research) according to the manufacturer's instructions. Bisulfite-converted DNA samples were randomly assigned to a chip well on the Infinium HumanMethylationEPIC BeadChip, amplified, hybridized onto the array, stained, washed, and imaged with the Illumina iScan SQ instrument to obtain raw image intensities.

Meffil R package (39) was used for the pre-processing of DNA methylation data. In the sample quality control, we removed the sex detection mismatches and the sex detection outliers (based on the difference between median chromosome Y and chromosome X probe intensities). We also discarded those samples whose predicted median methylated signal was more than 3 standard deviations from the expected. We excluded the outliers based on deviations from mean values for control probes (dye bias, bisulfite 1, and bisulfite 2). Finally, we removed those samples with more than 5% of undetected probes (detection *P*-value larger than 0.01) or with a low number of beads (less than 3). This quality control resulted in 3,424 individuals, indicating that 90,3% of the samples met our QC standards. In the feature quality control, we removed those probes undetected or with low bead numbers in more than 5% of the samples. We used InfiniumAnnotation (40) to filter probes where the 30bp 3'-subsequence of the probe is non-unique, probes with INDELs, probes with extension base inconsistent with specified color channel (type-I) or CpG (type-II) based on mapping, probes with a SNP in the extension base that causes a color channel switch from the official annotation, and probes where 5bp 3'-subsequence overlap with any of the SNPs with global population frequency higher than 1%. The functional normalization method was

further applied based on the first 10 principal components of the control probes. Consequently, the number of CpG probes kept was 745,150, which represents 86% of the total EPIC array manifest. CpG sites were annotated to genes using EPIC Illumina annotation ilm10b4.hg19. Blood cell type proportions were estimated using the blood gse35069 complete cell type methylation profile references from the *meffil* package. We then performed a surrogate variable analysis (SVA) to remove the batch effects using the *SmartSVA* package (41). Methylation levels were expressed as residual values after adjusting beta values for the first 60 surrogate variables.

Exposure and Clinical History Assessment

During the recruitment of participants, they were asked to complete a survey that included questions about personal information, medical history, social history, lifestyle, and family history. Alcohol and marijuana consumption was assessed by a 5-point scale ('never' to 'regularly'). Participants also reported their level of smoking according to 7 possible answers ('none' to 'more than 20 cigarettes per day'). Regarding the medical history, the survey covered information about the blood type, medications and supplements, and diagnosis of any type of disease (cardiovascular, respiratory, skin and hair, endocrine, gastrointestinal, genitourinary, musculoskeletal, neuropsychological, reproductive, immune, and cancer).

Statistical analyses and reproducibility

Epigenome-wide association analysis

The epigenome-wide association study (EWAS) was performed using the *MEAL* Bioconductor package (42). We performed a differential mean analysis on different substance consumption (tobacco, alcohol, and

marijuana) using the function *runPipeline* that calls *limma* (43). Based on a priori knowledge, we adjusted all the regression models by sex, age, ethnicity, body mass index (BMI), level of education, depression, anxiety, slide, cell type, and 60 surrogate variables. For each substance, we fitted models

$$E_j = \alpha_j + \beta_j S_j + \Sigma_r \gamma_r C_r + \varepsilon_j \qquad (1)$$

where E_j denotes the methylation level vector across individuals at probe j (j=1, ... 745150), S is the individuals' consumption (separated models for alcohol, smoking, and marijuana where fitted) with its associated effect, β_j, C_r is the r adjusting covariate and its effect γ_r, and ε_j is the noise that follows the distribution of methylation levels with mean 0. *P*-values obtained from each model were corrected for multiple comparisons using Bonferroni's correction. The inflation or deflation of *P*-values across the methylome was assessed with Q-Q plots and lambda values (44).

Enrichment analysis

Using the CpG sites with a *P*-value lower than $1 \cdot 10^{-4}$ from the EWAS results, we performed an enrichment in Kyoto Encyclopedia of Genes and Genomes (KEGG) and Gene Ontology (GO) pathways (45,46). We used the *gometh* function from the *missMethyl* BioConductor package (47). We also evaluated the over-representation of diseases using the DisGeNET platform that contains 1,134,942 gene-disease associations (48) and the *enricher* function from the *clusterProfiler* Bioconductor package (47).

Mediation analysis

We first used a multivariate mediation (*MultiMed* package (49)) for selecting the potential CpG sites mediating the effect of substance consumption on phenotypes. Second, we used the *mediate* function from

the *mediation* package (50) for estimating the single mediation by the most significant CpG sites found in the multivariate mediation step. To this end, we tested the total effect, the effect of the independent variable (substance consumption) on the mediation (CpG methylation), and the simultaneous effect of the mediator and the independent variable on the dependent variable (phenotype). Finally, we performed a causal mediation analysis, and we estimated the average causal mediation effects (ACME), the average direct effects (ADE), the total effect of the independent variable on the dependent variable, and the proportion of the effect of the independent variable on the dependent variable that goes through the mediator. We adjusted all the models by the same covariates as in the EWAS.

Results

We analyzed 3,424 individuals from the TruDiagnostic DNA Biobank recruited from the general population in EEUU. Table 1 presents demographic, and substance use characteristics. The mean age was 52.9 years (range: 13.3 – 97.8) and 58.7% were male. The participants were classified according to 7 ethnic groups and 'other' for those who had a mixed ethnicity. Most participants were Europeans (75.5%), and Latino American was the second most common ethnicity (8.1%). There were 149 current tobacco smokers, classified into seven groups according to the number of cigarettes smoked, and 3,275 non-smokers. Regarding alcohol consumption, there were 2,790 drinkers grouped by consumption frequency and 634 non-drinkers. Marijuana consumption was also classified according to the consumption frequency. In total, 2,908 did not smoke marijuana, and 367 smoke marijuana at least on special occasions.

Genome-wide effect of tobacco smoking on DNA methylation

We tested the association between the level of smoking (codified as a 7-point scale from 'none' to 'more than 20 cigarettes per day'; see Table 1) with each CpG site using linear regression models run in the *MEAL* R package (42). We found 528 CpG sites associated with smoking levels after correcting by multiple comparisons and genomic inflation was not observed ($\lambda = 1.031$). Table 2 lists the top 15 CpG sites for tobacco smoking (see Additional File 1: Table S1 for all the significant CpG sites). Fig. 1A shows how the CpG sites are distributed in the genome using a Manhattan plot. Among tobacco-related methylation sites, 68.2% were hypomethylated (that is, lower DNA methylation associated with higher tobacco consumption). If we consider only the top 15 CpG sites, a higher proportion of probes were hypomethylated (78.6%). From the 528 probes differentially methylated, 374 CpG mapped to 344 unique genes. *AHRR, GFI1, PRSS23,* and *IMMP2L* had 10, 6, 4, and 4 probes differentially methylated, respectively. Moreover, the differentially methylated genes with a *P*-value lower than $1 \cdot 10^{-4}$ were enriched in histone deacetylase complex, heterotrimeric G-protein complex, and GTPase complex GO pathways (Additional File 1: Table S2). The last two complexes are both involved in signal transduction. The most significant pathways in KEGG enrichment were alcoholism, morphine addiction, and dopaminergic and serotonergic synapses (Additional File 1: Table S3).

Consistent with previous studies, cg05575921 was the top-ranked CpG with a *P*-value = $1.3 \cdot 10^{-226}$. We further demonstrated that the effect of tobacco in this CpG site was proportional to the number of cigarettes smoked (Fig. 2A). Additionally, we compared our results with the ones previously reported in the EWAS catalog (51). This catalog contains all the associations CpG-trait with a *P*-value lower than $1 \cdot 10^{-4}$. In the case of the "smoking" trait, we were able to compare our results to 30 publications that

Table 2 Top 15 differentially methylated CpG sites by tobacco consumption.

CpG	chr	position	Gene Symbol	Gene Group	Less than 1 per week	Less than 1 per day	1-5 per day	6-10 per day	11-20 per day	More than 20 per day	F	P-Value	Adjusted P-Value	EWAS catalog
cg05575921	chr5	373378	*AHRR*	Body	-0.028	-0.054	-0.097	-0.216	-0.183	-0.217	208.2	1.3E-226	9.7E-221	Validated
cg21566642	chr2	233284661			-0.030	-0.058	-0.066	-0.109	-0.111	-0.160	59.7	1.6E-70	5.9E-65	Validated
cg01940273	chr2	233284934			-0.016	-0.034	-0.045	-0.066	-0.069	-0.109	53.6	2.4E-63	5.9E-58	Validated
cg03636183	chr19	17000585	*F2RL3*	Body	-0.007	-0.017	-0.029	-0.064	-0.053	-0.094	50.9	2.8E-60	5.2E-55	Validated
ch.6.276862 3F	chr6	144037285	*PHACT R2*	Body	-0.002	-0.001	0.000	0.009	-0.005	0.350	49.1	4.7E-58	7.0E-53	New
cg21161138	chr5	399360	*AHRR*	Body	-0.008	-0.016	-0.015	-0.071	-0.043	-0.069	40.6	5.7E-48	7.1E-43	Validated
cg17739917	chr17	38477572	*RARA*	5'UTR	-0.008	-0.053	-0.034	-0.069	-0.082	-0.116	38.2	4.5E-45	4.8E-40	Validated
cg15928392	chr2	81694446			0.000	0.000	0.001	-0.001	-0.002	-0.106	34.1	3.4E-40	3.2E-35	New
cg16774290	chr7	4824628	*AP5Z1*	Body	0.000	0.000	0.000	-0.001	0.000	-0.100	33.2	4.1E-39	3.4E-34	New
cg11173636	chr10	65632259			-0.008	0.017	0.024	0.018	0.339	-0.020	30.5	8.5E-36	6.3E-31	New
cg09935388	chr1	92947588	*GFI1*	Body	-0.018	-0.030	-0.029	-0.100	-0.069	-0.134	30.0	3.5E-35	2.4E-30	Validated
cg26703534	chr5	377358	*AHRR*	Body	-0.004	-0.013	-0.016	-0.052	-0.034	-0.049	29.9	4.0E-35	2.4E-30	Validated
cg14990808	chr6	28493651	*GPX5*	TSS200	-0.002	-0.022	0.005	0.002	0.001	0.246	26.3	1.1E-30	6.2E-26	New
cg25648203	chr5	395444	*AHRR*	Body	-0.006	-0.005	-0.011	-0.047	-0.037	-0.057	24.8	6.2E-29	3.3E-24	Validated
cg04176674	chr14	21121564			-0.004	-0.002	0.001	0.005	0.000	0.343	24.7	7.8E-29	3.9E-24	New

The CpG sites are annotated based on the chromosome (chr), the position (pos), the gene symbol from HGNC, and the gene group (based on the position of the CpG regarding the nearest gene). Each CpG site has a beta value for each consumer group vs non-consumers, a F-statistic (F), a nominal *p-value*, and an adjusted *p-value* by Bonferroni. The last column shows whether the CpG site has been previously reported in the EWAS catalog (validated) or not (new).

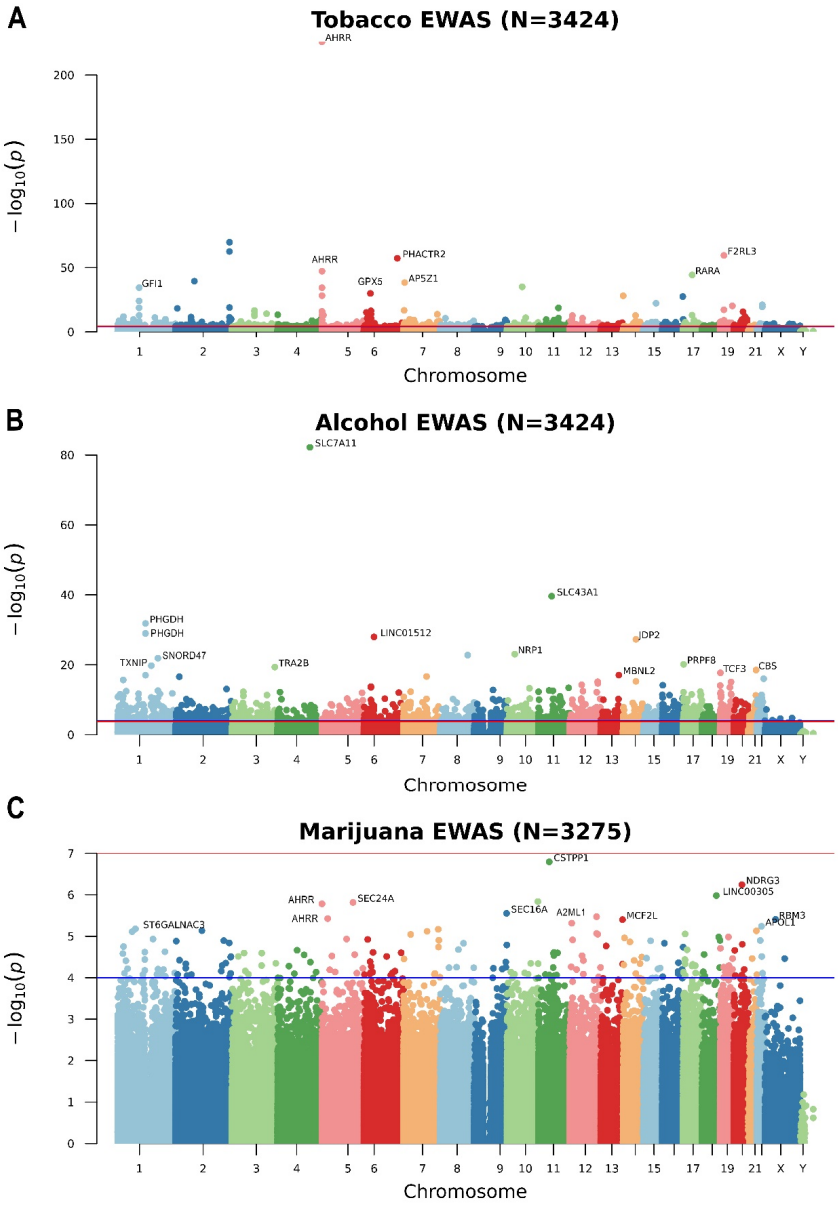


Fig. 1 Manhattan plots of the epigenome-wide association study (EWAS) of tobacco (**A**), alcohol (**B**), and marijuana (**C**) consumption. The Y-axis represents the -log10(p) values and the X-axis the position of the CpG sites within the chromosome. The blue line is the suggestive nominal *P*-value threshold (0.0001) and the red line is the *P*-value adjusted threshold lower than 0.05.

accumulated 19,156 significant CpG sites (P-value < $1 \cdot 10^{-4}$). From the 528 CpG sites differentially methylated in our analysis, we replicated 196 from the EWAS catalog and we identified 332 additional sites. More interestingly, from the top 50 probes in the EWAS catalog, we were able to replicate 46 with a P-value adjusted lower than 0.05 (Additional File 1: Table S4). Among the 332 new tobacco-related sites, 6 of them were in the top 15 CpG sites in our data, evidencing that tobacco may have important effects on them (Table 1).

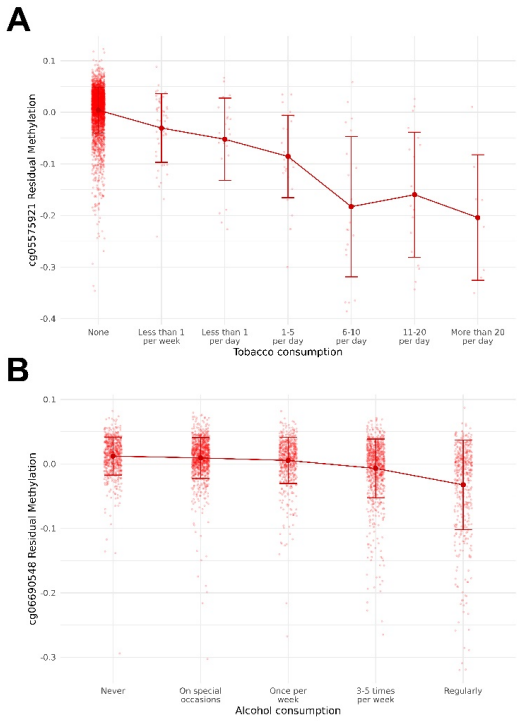


Fig. 2 Boxplots showing the association between CpG methylation and substance consumption. (**A**) Association between cg05575921 methylation (*AHRR*) and tobacco consumption. (**B**) Association between cg06690548 methylation (*SLC7A11*) and alcohol consumption. The Y-axis represents the residuals for beta values after adjusting by covariates. The X-axis represents the number of cigarettes smoked and the frequency of drinking, respectively. Methylation means for each tobacco consumption level are represented with their 95% confidence intervals.

Genome-wide effect of alcohol consumption on DNA methylation

We identified 2,569 CpG sites differentially methylated according to alcohol consumption frequency (5 levels from 'never' to 'regularly'; see Table 1). Model fitting showed no indication of genomic inflation (λ = 1.044). The top 15 CpG sites are shown in Table 3 and all the epigenome wide significant CpG sites are listed in Additional File 1: Table S5 and represented as a Manhattan Plot in Fig. 1B. Among them, 36.9% were hypomethylated for regular consumers compared with non-consumers. On the opposite, 12 out of the top 15 CpG sites (80%) were hypomethylated. Among the 2,569 alcohol-related methylation sites, 609 were intergenic and 1,960 were annotated to 1,670 unique genes. Five genes had seven or more significant probes mapping to their locus, including *RPTOR* (11 probes), *JARID2* (8), and *ABCG1* (8). The enrichment revealed an over-representation of different metabolic processes, such as vitamin B6, L-serine, and pyridoxal phosphate metabolisms (Additional File 1: Table S6). Among the enriched KEGG pathways, the vitamin B6 metabolism was again significantly enriched, along with the spliceosome, the cellular senescence, and the longevity regulating pathway (Additional File 1: Table S7). Additionally, the DisGeNET database revealed three interesting diseases associated with the genes differentially methylated: autistic disorder, acquired scoliosis, and curvature of the spine (Additional File 2: Fig. S1). The top CpG site (cg06690548, *P*-value = $5.9 \cdot 10^{-83}$) mapped to the *SLC7A11* gene and its methylation was significantly reduced proportionally to the alcohol consumption (Fig. 2B). In the EWAS catalog, there were 7,595 CpG sites associated with alcohol consumption with a *P*-value lower than $1 \cdot 10^{-4}$ based on 6 publications available. We replicated 286 CpG sites and we identified 2,283 new alcohol-related sites with a *P*-value adjusted lower than 0.05 (3 of them among our top sites). Additionally, 33 out of the top 50 probes in the EWAS catalog were

replicated in our data with a *P*-value adjusted lower than 0.05 (Additional File 1: Table S8).

Genome-wide effect of marijuana consumption on DNA methylation

In the EWAS for the frequency of marijuana use (codified as a 5-point scale from 'never' to 'regularly'; see Table 1), we did not find any CpG site with a *P*-value adjusted lower than 0.05 (Additional File 1: Table S9 and Fig. 1C). However, we found 195 CpG sites at a suggestive significant level (*P*-value < $1 \cdot 10^{-4}$). From the top 15 CpG sites, 9 CpG sites were hypomethylated for regular consumption compared to no consumption (Table 4). Gene symbols for the 195 differentially methylated CpG sites at suggestive *P*-value were tested for enrichment in KEGG pathways and Gene Ontology (GO). There was strong evidence for enrichment of many GO terms related to the neurotoxic effect of marijuana consumption (Additional File 1: Table S10). The strongest associations included paranodal junction assembly (*P*-value = $6.6 \cdot 10^{-5}$), myelin assembly (*P*-value = $7.1 \cdot 10^{-5}$), and neuromuscular process controlling balance (*P*-value = $2.3 \cdot 10^{-4}$). We also looked at the EWAS catalog and we only found 1 publication available for lifetime cannabis use (ever vs never) (29). In that publication, they found 118 CpG sites with a *P*-value lower than $1 \cdot 10^{-4}$. Comparing those sites with the 195 CpG sites from our analysis, we did not find overlapping (Additional File 1: Table S11). This may be in part due to the differences in the variable of interest and the study population, such as the evaluation of the effect of cannabis use on non-Hispanic white women and the risk of breast cancer (29).

Comparison between tobacco, alcohol, and marijuana effects

We further compared the genes differentially methylated by tobacco, alcohol, and marijuana consumption. We selected the CpG sites

Table 3 Top 15 differentially methylated CpG sites by alcohol consumption.

CpG	chr	position	Gene Symbol	Gene Group	On special occasions	Once per week	3-5 times per week	Regularly	F	P-Value	Adjusted P-Value	EWAS catalog
cg06690548	chr4	139162808	*SLC7A11*	Body	-0.003	-0.005	-0.015	-0.031	103.1	5.9E-83	4.4E-77	Validated
cg11376147	chr11	57261198	*SLC43A1*	Body	-0.001	-0.002	-0.006	-0.010	49.3	2.4E-40	8.9E-35	Validated
cg26457483	chr1	120256112	*PHGDH*	Body	-0.001	-0.004	-0.017	-0.028	39.7	1.5E-32	3.8E-27	Validated
cg14476101	chr1	120255992	*PHGDH*	Body	-0.002	-0.005	-0.016	-0.031	36.2	1.1E-29	2.0E-24	Validated
cg18120259	chr6	43894639	*LINC01512*	Body	-0.002	-0.004	-0.009	-0.019	35.0	1.1E-28	1.6E-23	Validated
cg06088069	chr14	75895604	*JDP2*	5'UTR	-0.002	-0.003	-0.007	-0.013	34.2	5.0E-28	6.1E-23	Validated
cg21998542	chr10	33605101	*NRP1*	Body	-0.002	-0.005	-0.009	-0.019	29.0	9.2E-24	9.8E-19	Validated
cg15837522	chr8	117892654			-0.003	-0.011	-0.022	-0.031	28.7	1.7E-23	1.6E-18	Validated
cg06644515	chr1	173834831	*SNORD47;GAS5; SNORD81;SNORD80; SNORD78;SNORD79*	TSS1500; Body; TSS1500; TSS1500; TSS200;TSS1500	-0.001	-0.001	-0.007	-0.013	27.6	1.3E-22	1.1E-17	Validated
cg12116137	chr17	1576449	*PRPF8*	Body	0.001	0.001	0.008	0.018	25.5	7.5E-21	5.5E-16	New
cg19693031	chr1	145441552	*TXNIP*	3'UTR	-0.004	-0.003	-0.011	-0.026	25.1	1.8E-20	1.2E-15	Validated
cg12825509	chr3	185648568	*TRA2B*	Body	0.000	-0.004	-0.010	-0.016	24.6	4.5E-20	2.8E-15	Validated
cg14346162	chr21	44490229	*CBS*	Body	0.001	0.004	0.009	0.014	23.5	3.1E-19	1.8E-14	New
cg12973487	chr19	1623075	*TCF3*	Body	0.004	0.005	0.007	0.014	22.6	2.0E-18	1.0E-13	Validated
cg05713943	chr13	97912352	*MBNL2*	5'UTR	0.001	0.000	-0.005	-0.008	21.8	8.7E-18	4.2E-13	New

The CpG sites are annotated based on the chromosome (chr), the position (pos), the gene symbol from HGNC, and the gene group (based on the position of the CpG regarding the nearest gene). Each CpG site has a beta value for each consumer group vs non-consumers, a F-statistic (F), a nominal *p-value*, and an adjusted *p-value* by Bonferroni. The last column shows whether the CpG site has been previously reported in the EWAS catalog (validated) or not (new).

Table 4 Top 15 differentially methylated CpG sites by marijuana consumption.

CpG	chr	position	Gene Symbol	Gene Group	On special occasions	Once per week	3-5 times per week	Regularly	F	P-Value	Adjusted P-Value
cg05107281	chr11	47072710	*CSTPP1*	Body	-0.004	-0.010	-0.014	0.001	9.4	1.6E-07	0.12
cg06499565	chr20	35374747	*NDRG3*	TSS1500	0.001	0.006	0.000	-0.001	8.7	5.7E-07	0.17
cg10054857	chr18	61816543	*LINC00305*	TSS1500	-0.004	-0.007	-0.002	-0.002	8.4	1.0E-06	0.17
cg24344693	chr10	133273964			-0.004	0.003	-0.001	-0.002	8.2	1.5E-06	0.17
cg20226924	chr5	133985272	*SEC24A*	Body	0.001	0.003	0.003	0.001	8.2	1.5E-06	0.17
cg05575921	chr5	373378	*AHRR*	Body	-0.003	-0.011	-0.012	-0.020	8.1	1.7E-06	0.17
cg21161138	chr5	399360	*AHRR*	Body	-0.004	-0.011	-0.009	-0.012	8.1	1.7E-06	0.17
cg19730404	chr9	139361517	*SEC16A; SEC16A*	ExonBnd;Body	-0.003	-0.001	0.001	-0.004	7.8	2.8E-06	0.25
cg19308363	chr12	116290566			0.006	0.005	-0.005	-0.015	7.7	3.4E-06	0.25
cg08258765	chr5	24841586	*LOC340107*	TSS1500	0.005	0.002	0.004	-0.002	7.7	3.8E-06	0.25
cg03838168	chrX	48433876	*RBM3*	Body	-0.003	-0.004	0.015	0.005	7.7	3.9E-06	0.25
cg16822035	chr13	113633379	*MCF2L;MCF2L*	Body;TSS1500	0.001	0.004	0.009	0.012	7.6	4.0E-06	0.25
cg11756734	chr12	9028945	*A2ML1*	3'UTR	-0.005	0.002	-0.007	0.003	7.5	4.8E-06	0.26
cg08415592	chr22	36648973	*APOL1*	TSS200	0.008	0.016	-0.004	0.007	7.4	5.8E-06	0.26
cg17325792	chr1	77042560	*ST6GALNAC3; ST6GALNAC3*	3'UTR;Body	0.003	0.001	0.003	-0.011	7.4	6.6E-06	0.26

The CpG sites are annotated based on the chromosome (chr), the position (pos), the gene symbol from HGNC, and the gene group (based on the position of the CpG regarding the nearest gene). Each CpG site has a beta value for each consumer group vs non-consumers, a F-statistic (F), a nominal *p-value*, and an adjusted *p-value* by Bonferroni.

differentially methylated with a *P*-value lower than $1 \cdot 10^{-5}$ and extracted the genes where they were annotated. We found 25 genes associated with more than one drug (Fig. 3). Using the GeneMANIA bioinformatics software (52), which contains information on genetic and physical interactions, shared protein domains, and co-expression data, we found that 23 out of the 25 genes were connected to form a compact cluster network (Additional File 2: Fig. S2). This cluster was enriched in macromolecule and protein deacetylation and transcription regulator complex functions. The largest overlap was between alcohol and tobacco, revealing 20 genes differentially methylated by these two drugs independently. Among these genes, *NOP53, ZFP36L1, PTK2,* and *PPP1R16B* were associated with phosphatidylinositol-mediated signaling and inositol lipid-mediated signaling, two signaling events present in the nervous system (53). In addition, *IFT140, HDAC4, PPP1R16B*, and *DMTN* were associated with cell projection assembly. Alcohol and marijuana overlapped in 3 genes, being *CUX1* the most interesting due to its relationship with neuronal differentiation in the brain. Finally, two genes were differentially methylated due to tobacco and marijuana use (*AHRR* and *LOC440839*). It is worth mentioning that the CpG sites mapped to the *AHRR* gene are the same in tobacco and marijuana (cg05575921 and cg21161138) and it is present in the top 15 probes for both EWAS.

Mediation between substance consumption and hypertension by CpG methylation

We evaluated whether the changes at the CpG methylation level mediated the effect of drugs on hypertension. We first tested the association between smoking and hypertension. We considered that the group with the highest levels of smoking were those who smoked more than 11 cigarettes per day, joining the categories 11-20 and >20 cigarettes due to their low numbers. We evaluated the association between smoking codified as numeric and

high blood pressure and we did not detect a significant tendency (*P*-value = 0.26). The forest plot shows that the risk of hypertension increases with a higher number of cigarettes smoked except for the last group (more than 11 cigarettes per day) (Additional File 2: Fig. S3). Although we expected this group to be the one at the highest risk, we also observed that those participants were also the youngest (average of 5.75 years less, *P*-value = 0.005), suggesting a particularly strong healthy donor effect for this group. We tested the association after removing this group and we found a significant association between tobacco smoking and hypertension (*P*-value = 0.009, OR = 1.28). The forest plot in Fig. 4A revealed a clear dose-response relationship where individuals who consume 6 to 10 cigarettes per day have 3.19 times of high blood pressure risk compared with non-smokers (P-value = 0.023). We used the top 200-ranked CpG sites in a multivariate mediation analysis between tobacco smoking and hypertension and the *MultiMed* package (49) revealed no CpG sites with a *P*-value lower than 0.05 (Additional File 1: Table S12).

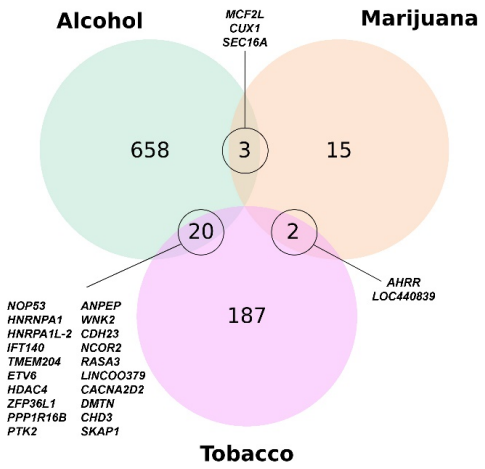


Fig. 3 Venn diagram comparing genes differentially methylated for tobacco (pink), alcohol (green), and marijuana (orange) consumption. The overlapped genes are annotated.

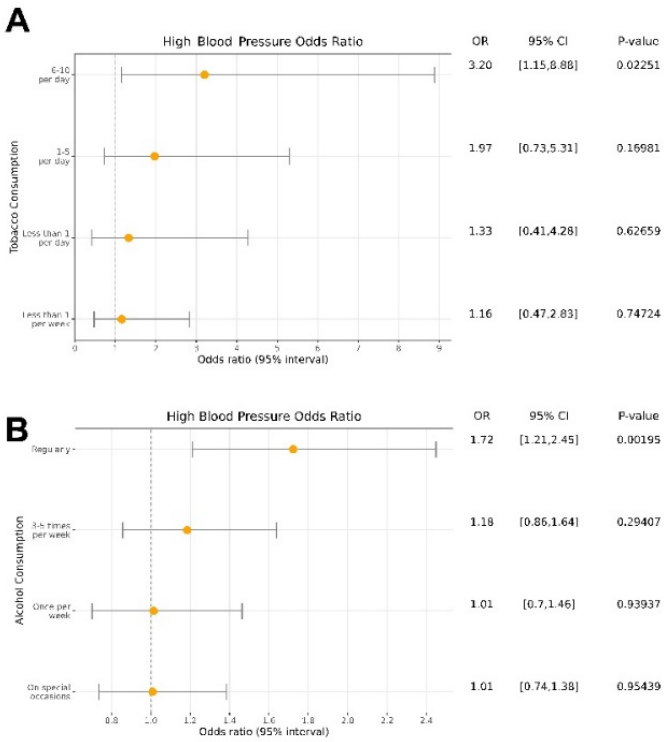


Fig. 4 Forest plot of the association between tobacco (**A**) and alcohol (**B**) consumption with hypertension. OR: Odds Ratio.

We also tested the association between marijuana consumption and high blood pressure. We did not find any significant association when comparing the 4 levels of consumption (from 'on special occasions' to 'regularly') with non-consumers (Additional File 2: Fig. S4).

As for alcohol use, we found a significant association between alcohol and higher blood pressure (*P*-value = 0.001, OR = 1.13). In addition, the forest plot in Fig. 4B revealed a significant association between daily consumption with high blood pressure (*P*-value = 0.014, OR = 1.39) and a non-significant association between light to moderate consumption with the phenotype, as expected. Thus, using the top 200 CpG sites from the alcohol

EWAS, we performed a multivariate mediation analysis between alcohol consumption and hypertension. We found 8 significant mediators that coincide with the top 6 CpG sites and two other CpG sites in the EWAS analysis (Table 5 and Additional File 1: Table S13). To see the effect of each CpG site, we performed a univariate mediation analysis for each of these CpG sites (Table 5). All the CpG sites except one mediated independently between 12.5 and 30.9% of the effect of alcohol on hypertension. The most differentially methylated CpG site by alcohol consumption, cg06690548, was also the most significant mediator between alcohol consumption and high blood pressure after adjusting by covariates. This CpG site mediated 73.6% of the total effect of alcohol on the phenotype (*P*-value = 0.008). As mentioned before, heavy alcohol consumption increases by 39% the risk of hypertension. With our data, we have demonstrated that this association is mostly due to changes in DNA methylation.

Table 5 Mediation analysis between alcohol consumption and hypertension using methylation level of CpG sites as mediators.

CpG	Multivariate Analysis		Univariate Analysis					
			Mediated effect		Direct effect		Proportion mediated	
	S-statistic	*P*-Value	Estimate	*P*-value	Estimate	*P*-value	Estimate	*P*-value
cg06690548	0.0344	0	0.0088	0	0.0032	0.502	0.736	0.008
cg11376147	0.0146	0.004	0.0034	0	0.0085	0.05	0.282	0.004
cg26457483	0.0146	0.004	0.0037	0	0.0087	0.062	0.297	0.008
cg14476101	0.0105	0.04	0.0028	0.002	0.0091	0.044	0.227	0.02
cg18120259	0.0161	0.004	0.0038	0	0.0084	0.07	0.309	0.014
cg06088069	0.0124	0.014	0.0028	0	0.0093	0.048	0.231	0.008
cg16740586	0.0105	0.04	0.0016	0.01	0.0108	0.016	0.125	0.016
cg15659943	0.0124	0.014	0.0028	0	0.0093	0.036	0.228	0.004

Discussion

The current study evaluated the effect of tobacco, alcohol, and marijuana consumption on genome-wide DNA methylation in 3,424 individuals from an EEUU population-based cohort. We identified 528 CpG sites differentially methylated according to tobacco smoking, 2,569 according to alcohol consumption, and 195 suggestive associations of marijuana consumption (Nominal *P*-value < $1 \cdot 10^{-4}$). Second, we detected a large overlapping between the differentially methylated genes by these three unhealthy lifestyle habits. Third, we found a significant mediation between alcohol consumption and hypertension by the top alcohol-related methylation sites.

A considerable amount of literature has been published on DNA methylation changes due to smoking. The first studies evaluating these changes were carried out in small panels of genes (54,55). It was not until 2012 that the first epigenome-wide association study on tobacco was reported by Wan et al. (56). Since that time, several studies have demonstrated the huge impact of tobacco on DNA methylation across the human genome, even in newborns when the tobacco exposure was during pregnancy (15,16,18–20,57). Our results in TruDiagnostic DNA Biobank replicated previous studies revealing a high impact of smoking on DNA methylation epigenome-wide. The EWAS catalog (51) is the biggest database of epigenome-wide association studies containing associations between CpG sites and traits with a *P*-value lower than $1 \cdot 10^{-4}$. In the case of tobacco smoking, this catalog gathers significant associations from 30 publications. From the top 50 probes in the EWAS catalog, we replicated 46. We also identified 332 tobacco-related CpG sites that were not previously reported in the EWAS catalog. Among them, it is remarkable the CpG at *PHACTR2* gene because it is involved in actin cytoskeleton organization and implicated in Parkinson's disease (58), and the CpG at

GPX5 gene since it protects cells and enzymes from oxidative damage. Additionally, our results confirm previous observations where the cg05575921 mapped to *AHRR* (*P*-value = $9.7 \cdot 10^{-221}$) and the cg21566642 in the 2q37.1 region (*P*-value = $5.9 \cdot 10^{-65}$) were the most significantly associated CpG sites to tobacco consumption (19–21).

Alcohol is known to affect DNA methylation. To date, several EWAS have detected CpG sites associated with alcohol consumption (22,24–26,59,60). Here, we identified more than 2,500 CpG sites associated with drinking. Interestingly, these genes were highly associated with autistic disorder, acquired scoliosis, and curvature of the spine. Fetal alcohol spectrum disorder (FASD) is the general term that encompasses the range of adverse effects associated with alcohol exposure during pregnancy. Among FASD symptoms, it is common to find autistic-like traits, as well as scoliosis and other musculoskeletal anomalies (61–63). In addition, previous studies have already investigated the epigenetic mechanism linking autism and FASD (64). The EWAS catalog collects information from 6 publications and our results replicated 33 out of the 50 top CpG sites. As we identified, many studies detected cg06690548 mapped at the *SLC7A11* promoter as the most alcohol-related methylation site (22,24,25). Furthermore, Lohoff et al. demonstrated that various liver biomarkers were robustly associated with *SLC7A11* methylation status (24), suggesting an implication of this gene in the disturbance of the gastrointestinal system when consuming alcohol. Further, we identified many CpG sites that were not reported in the EWAS catalog. Among them, three sites at *PRPF8, CBS*, and *MBNL2* genes, respectively, were in the top-ranked differentially methylated probes in our data. *PRPF8* and *MBNL2* are involved in pre-mRNA alternative splicing regulation, and *CBS* is translated to the cystathionine beta-synthase enzyme.

The studies that have investigated DNA methylation modifications after marijuana consumption are limited. In 2015, Watson et al. evaluated in rats the effect of cannabis parental exposure on the epigenome of the nucleus accumbens (65) and they identified 1027 differentially methylated regions. Five years later, Osborne et al. carried out the first EWAS on heavy cannabis consumption with and without tobacco comparing 48 consumers with 48 controls (28). They found five differentially methylated sites in cannabis and tobacco users that replicated previous studies on the effects of tobacco. However, cannabis-only users had no evidence of significant differential methylation in any gene. Markunas et al. performed another EWAS with a larger sample size (1,247 ever users) consisting of women at risk of developing breast cancer (29). They identified a unique significant CpG mapped to *CEMIP* 5' region. However, they designed a biomarker for lifetime cannabis use based on the top 50 EWAS CpG sites. In our study, 367 individuals smoked from occasionally to daily. The EWAS did not reveal significant CpG sites at the Bonferroni adjustment. Nonetheless, the 195 CpG sites with a *P*-value lower than $1 \cdot 10^{-4}$ were enriched in genes related to paranodal junction assembly, myelin assembly, and neuromuscular process controlling balance. This suggests a possible implication of DNA methylation changes on the long-term neurotoxic effects of marijuana smoking. In addition, we detected cg05575921 (*AHRR*), the most significant tobacco-associated CpG site, differentially methylated according to marijuana use with a nominal *P*-value equal to $1.7 \cdot 10^{-6}$. Allen et al. already found that the link between marijuana use and epigenetic age acceleration was statistically mediated via hypomethylation at site cg05575921 (66). This is consistent with the association of the *AHRR* gene with exposure to tobacco and fine particulate matter (PM2.5) which suggests that marijuana inhalation can produce similar effects (66,67).

In the current study, we compared the genes differentially methylated by tobacco, alcohol, and marijuana at a *P*-value nominal lower than $1 \cdot 10^{-5}$. Surprisingly, we found that 25 genes were affected by more than one drug and these genes formed a compact underlying genetic network. From them, 20 genes were differentially methylated by alcohol and tobacco consumption, and they were enriched in signaling pathways involved in neurons and cell projection assembly. Among the 3 genes that overlapped between alcohol and marijuana effects, it is remarkable the implication in neuronal differentiation. *AHRR* gene was involved in tobacco and marijuana epigenetic changes. As mentioned previously, it may be explained by the effects of smoke inhalation on this gene. This overlapping between different drugs suggests that similar molecular pathways are affected by similar unhealthy lifestyle habits. This finding may provide insight into new targets for treating addiction and for preventing common diseases associated with drug consumption, like cardiovascular disease.

In our data, tobacco was partially associated with hypertension, and alcohol consumption was highly associated with the condition, as demonstrated previously (32–35). Marijuana was not associated with hypertension, in line with previous studies that have revealed ambiguous results (36–38). In the case of tobacco exposure, we found unexpected results because the individuals who smoked the most were the ones who had less risk to develop hypertension. These results may be explained due to the lower age of the individuals in that group and also to the healthy donor effect of the data. This means that participants are volunteers who have paid for the TruAge test and may have healthy habits that protect them against hypertension although they are heavy smokers.

Our data replicated prior studies where light to moderate drinking was not associated with high blood pressure and heavy drinking increased the risk of the disease (5,33). Another important finding was that CpG methylation

significantly mediated the effect of alcohol consumption on hypertension. The multivariate mediation analysis revealed eight CpG sites as potential mediators that included the top six alcohol-related probes and were previously reported in the EWAS catalog. All of them were hypomethylated in heavy consumption, in line with one study that revealed a global hypomethylation in hypertensive patients (68). More interestingly, lower methylation levels of cg06690548 at *SLC7A11* and cg14476101 at *PHGDH* have been seen previously associated with higher systolic and diastolic blood pressure (69–71). Additionally, hypomethylation of cg06690548 and cg14476101 were associated with higher expression of *SLC7A11* and *PHGDH*, respectively (71). *SLC7A11* enhances antioxidant defense and protects against endothelial dysfunction and vascular inflammation. This increases vascular tone and rigidity, and consequently blood pressure. Also, Richard et al. evidenced triangular associations between methylation, gene expression, and blood pressure (70). We also tested univariate mediation based on cg06690548 methylation and we revealed that 73.6% of the effect of alcohol on high blood pressure was mediated by the CpG methylation level (*P*-value = 0.008). Besides, *PHGDH* encodes the enzyme which is involved in the early steps of serine synthesis, which is highly related to tissue growth. Thus, hypomethylation of this gene may act in the vascular adaptation to body-tissue growth during adolescence (71). In essence, we have demonstrated that the effect of heavy drinking on high blood pressure is partially mediated by hypomethylation of CpG sites that are significantly associated with the disease. This finding provides new insights on targets to prevent and manage hypertension in individuals with regular alcohol consumption.

The generalizability of these results is subject to certain limitations. First, DNA methylation was obtained from blood samples, thus, further research is required to understand the implication of the identified markers in each

tissue. Second, genetics has an important role in substance use predisposition. In our analysis, we were not able to remove the genetic factor because of the lack of data. Some of the differentially methylated probes may be a consequence of the genetic differences and not the exposure itself. Notwithstanding this limitation, we filtered all the probes with a SNP in the extension base and all probes where 5bp 3'-subsequence overlapped with any of the SNPs with a global population frequency higher than 1%. Third, the consumption assessment was self-reported and not specific for a time period, limiting the credibility. In addition, we did not have information on whether marijuana was smoked mixed or not with tobacco. This information could benefit future studies on removing the tobacco effect. Fourth, we have compared our results with the EWAS catalog, that do not contain all the published EWAS. However, this catalog serves as a reference since it contains a large number of published papers.

Our study also had notable strengths, including a large number of drinkers and the high variability in drinking frequency. This allowed us to test the mediation analysis between alcohol consumption and hypertension. Moreover, most studies are focused on evaluating the effects of one substance in drug-specific cohorts. Our data provided information on tobacco, alcohol, and marijuana consumption in the same individuals, along with clinical data. Since our results for tobacco and alcohol are comparable with previous studies, we may assume that our results for marijuana consumption are reliable, suggesting that marijuana does not have a big effect on blood DNA methylation.

Conclusions

To the best of our knowledge, this is the first study to assess simultaneously the effect of tobacco, alcohol, and marijuana on DNA methylation. We have shown that tobacco and alcohol have large effects on genome-wide DNA

methylation, while marijuana consumption has small effects. Most importantly, many genes differentially methylated by smoking are also affected by alcohol and marijuana consumption, suggesting a similar epigenetic impact after drug consumption. The results of this research also have significant implications for the understanding of how alcohol consumption increase hypertension. We demonstrated that the most alcohol-related CpG sites are important mediators of the effect of alcohol consumption on high blood pressure. Finally, the current data highlight the importance of investigating methylation biomarkers in blood to monitor diseases, such as neurological disorders derived from substance consumption.

Supplementary information

Additional file 1: Table S1. All epigenome-wide significant CpG sites for tobacco smoking (*P*-value adjusted < 0.05). Table S2. Top enriched gene ontology terms in tobacco EWAS. Table S3. Top enriched KEGG terms in tobacco EWAS. Table S4. Top CpG sites in the EWAS catalog for "smoking" trait. Table S5. All epigenome-wide significant CpG sites for alcohol consumption (*P*-value adjusted < 0.05). Table S6. Top enriched gene ontology terms in alcohol EWAS. Table S7. Top enriched KEGG terms in alcohol EWAS. Table S8. Top CpG sites in the EWAS catalog for "alcohol" trait. Table S9. All epigenome-wide significant CpG sites for marijuana consumption (*P*-value < 0.0001). Table S10. Top enriched gene ontology terms in marijuana EWAS. Table S11. Top CpG sites in the EWAS catalog for "lifetime cannabis use" trait. Table S12. Multivariate mediation results for the top 200 CpG sites in the tobacco EWAS. Table S13. Multivariate mediation results for the top 200 CpG sites in the alcohol EWAS.

Additional file 2: Figure S1. DisGeNET enrichment results in alcohol EWAS. Figure S2. Network image from GeneMANIA software based on the 25 genes overlapped between alcohol, tobacco, and marijuana consumption. Figure S3. Forest plot of the association between tobacco smoking and high blood pressure. Figure S4. Forest plot of the association between marijuana consumption and high blood pressure.

Acknowledgements

We are grateful to all participants and researchers who took part in this study.

Author contributions

JRG conceived the study and supervised analyses. JRG and NC-G designed the analysis. NC-G normalized data and performed the statistical analyses. VBD, RS, HW, and TLM generated and coordinated the data, secured ethics approval and consents. NC-G wrote the original draft of the paper and JRG, AC, VBD, and RS contributed to reviewing and editing the manuscript. All authors read and approved the final version of the manuscript.

Funding

This research has received funding from the Spanish Ministry of Science and Innovation through the “Centro de Excelencia Severo Ochoa 2019-2023 (CEX2018-000806-S) program, and support from the Generalitat de Catalunya through the CERCA Program. NC is supported by Spanish regional program PERIS (Ref.: SLT017/20/000061), granted by Departament de Salut de la Generalitat de Catalunya. TruDiagnostics also provided funding for data analysis.

Availability of data and materials

The data that support the findings of this study are available upon reasonable request due to ensuring privacy of the participants. Please e-mail varun@trudiagnostic.com for data requests. Any custom code or software used in our analysis is available at DOI: 10.5281/zenodo.6417926 (URL: https://zenodo.org/badge/latestdoi/296552532).

Declarations

Ethics approval and consent to participate

The study involving human participants was reviewed and approved by the IRCM IRB. The patients/participants provided their written informed consent to participate in this study.

Consent for publication

Not applicable.

Competing interests

NC has nothing to declare. VBD, RS, HW, and TLM are employees of TruDiagnostic, the company that has provided the data for this study. JRG has received funding from TruDiagnostic as a scientific advisor.

References

1. Substance Abuse and Mental Health Services Administration (SAMHSA). Results from the 2015 National Survey on Drug Use and Health: Summary of National Findings [Internet]. 2017 [cited 2022 Jun 27]. Available from: https://www.samhsa.gov/

2. World Health Organization. Management of Substance Abuse Team. The health and social effects of nonmedical cannabis use. 2016.

3. Geneva: World Health Organization. WHO global report on trends in prevalence of tobacco use 2000-2025, fourth edition. 2021.

4. West R. Tobacco smoking: Health impact, prevalence, correlates and interventions. Psychol Health [Internet]. 2017 Aug 3 [cited 2022 Jun 20];32(8):1018–36. Available from: https://pubmed.ncbi.nlm.nih.gov/28553727/

5. O'Keefe JH, Bhatti SK, Bajwa A, DiNicolantonio JJ, Lavie CJ. Alcohol and cardiovascular health: the dose makes the poison…or the remedy. Mayo Clin Proc [Internet]. 2014 [cited 2022 Jun 27];89(3):382–93. Available from: https://pubmed.ncbi.nlm.nih.gov/24582196/

6. Geneva: World Health Organization. Global status report on alcohol and health 2018. 2018.

7. Egger G, Liang G, Aparicio A, Jones PA. Epigenetics in human disease and prospects for epigenetic therapy. Nature [Internet]. 2004 May 27 [cited 2022 May 3];429(6990):457–63. Available from: https://pubmed.ncbi.nlm.nih.gov/15164071/

8. Szutorisz H, Hurd YL. Epigenetic Effects of Cannabis Exposure. Biol Psychiatry [Internet]. 2016 [cited 2022 May 4];79(7):586. Available from: /pmc/articles/PMC4789113/

9. Cecil CAM, Walton E, Viding E. DNA Methylation, Substance Use and Addiction: a Systematic Review of Recent Animal and Human Research from a Developmental Perspective. Curr Addict Reports [Internet]. 2015 Dec 1 [cited 2022 Jul 8];2(4):331–46. Available from: https://link.springer.com/article/10.1007/s40429-015-0072-9

10. Moosavi A, Ardekani AM. Role of Epigenetics in Biology and Human Diseases. Iran Biomed J [Internet]. 2016 Nov 1 [cited 2022 Jul 13];20(5):246. Available from: /pmc/articles/PMC5075137/

11. Leenen FAD, Muller CP, Turner JD. DNA methylation: conducting the orchestra from exposure to phenotype? Clin Epigenetics [Internet]. 2016 Sep 6 [cited 2022 May 2];8(1). Available from: https://pubmed.ncbi.nlm.nih.gov/27602172/

12. Carreras-Gallo N, Cáceres A, Balagué-Dobón L, Ruiz-Arenas C, Andrusaityte S, Carracedo Á, et al. The early-life exposome modulates the effect of polymorphic inversions on DNA methylation. Commun Biol [Internet]. 2022 Dec [cited 2022 Jul 20];5(1). Available from: https://pubmed.ncbi.nlm.nih.gov/35550596/

13. Li E, Zhang Y. DNA Methylation in Mammals. Cold Spring Harb Perspect Biol [Internet]. 2014 [cited 2022 Jul 19];6(5). Available from: /pmc/articles/PMC3996472/

14. Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. Genome Res [Internet]. 2010 Mar [cited 2022 Jul 20];20(3):320–31. Available from: https://pubmed.ncbi.nlm.nih.gov/20133333/

15. Dugué PA, Jung CH, Joo JE, Wang X, Wong EM, Makalic E, et al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. Epigenetics [Internet]. 2020 Apr 2 [cited 2022 Jun 2];15(4):358–68. Available from: https://pubmed.ncbi.nlm.nih.gov/31552803/

16. Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One [Internet]. 2013 May 17 [cited 2022 Jun 2];8(5). Available from: https://pubmed.ncbi.nlm.nih.gov/23691101/

17. Dogan M V., Beach SRH, Philibert RA. Genetically contextual effects of smoking on genome wide DNA methylation. Am J Med Genet B Neuropsychiatr Genet [Internet]. 2017 Sep 1 [cited 2022 Jun 2];174(6):595–607. Available from: https://pubmed.ncbi.nlm.nih.gov/28686328/

18. Christiansen C, Castillo-Fernandez JE, Domingo-Relloso A, Zhao W, El-Sayed Moustafa JS, Tsai PC, et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clin Epigenetics [Internet]. 2021 Dec 1 [cited 2022 Jun 2];13(1). Available from: https://pubmed.ncbi.nlm.nih.gov/33593402/

19. Sikdar S, Joehanes R, Joubert BR, Xu CJ, Vives-Usano M, Rezwan FI, et al. Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. Epigenomics [Internet]. 2019 Oct 1 [cited 2022 Jun 2];11(13):1487–500. Available from: https://pubmed.ncbi.nlm.nih.gov/31536415/

20. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect [Internet]. 2012 Oct [cited 2022 Jul 7];120(10):1425–31. Available from: https://pubmed.ncbi.nlm.nih.gov/22851337/

21. Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet [Internet]. 2015 Apr 15 [cited 2022 Jul 7];24(8):2201–17. Available from: https://pubmed.ncbi.nlm.nih.gov/25552657/

22. Xu K, Montalvo-Ortiz JL, Zhang X, Southwick SM, Krystal JH, Pietrzak RH, et al. Epigenome-Wide DNA Methylation Association Analysis Identified Novel Loci in Peripheral Cells for Alcohol Consumption Among European American Male Veterans. Alcohol Clin Exp Res [Internet]. 2019 Oct 1 [cited 2022 May 12];43(10):2111–21. Available from: https://pubmed.ncbi.nlm.nih.gov/31386212/

23. Toinét Cronjé H, Elliott HR, Nienaber-Rousseau C, Pieters M. Replication and expansion of epigenome-wide association literature in a black South African population. Clin Epigenetics [Internet]. 2020 Jan 7 [cited 2022 May 12];12(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31910897/

24. Lohoff FW, Clarke TK, Kaminsky ZA, Walker RM, Bermingham ML, Jung J, et al. Epigenome-wide association study of alcohol consumption in N = 8161 individuals and relevance to alcohol use disorder pathophysiology: identification

of the cystine/glutamate transporter SLC7A11 as a top target. Mol Psychiatry [Internet]. 2022 [cited 2022 May 18];27(3). Available from: https://pubmed.ncbi.nlm.nih.gov/34857913/

25. Dugué PA, Wilson R, Lehne B, Jayasekara H, Wang X, Jung CH, et al. Alcohol consumption is associated with widespread changes in blood DNA methylation: Analysis of cross-sectional and longitudinal data. Addict Biol [Internet]. 2021 Jan 1 [cited 2022 May 5];26(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31789449/

26. Wilson LE, Xu Z, Harlid S, White AJ, Troester MA, Sandler DP, et al. Alcohol and DNA Methylation: An Epigenome-Wide Association Study in Blood and Normal Breast Tissue. Am J Epidemiol [Internet]. 2019 Jun 1 [cited 2022 May 12];188(6):1055–65. Available from: https://pubmed.ncbi.nlm.nih.gov/30938765/

27. Stephenson M, Bollepalli S, Cazaly E, Salvatore JE, Barr P, Rose RJ, et al. Associations of Alcohol Consumption With Epigenome-Wide DNA Methylation and Epigenetic Age Acceleration: Individual-Level and Co-twin Comparison Analyses. Alcohol Clin Exp Res [Internet]. 2021 Feb 1 [cited 2022 May 12];45(2):318–28. Available from: https://pubmed.ncbi.nlm.nih.gov/33277923/

28. Osborne AJ, Pearson JF, Noble AJ, Gemmell NJ, Horwood LJ, Boden JM, et al. Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort. Transl Psychiatry 2020 101 [Internet]. 2020 Apr 22 [cited 2022 Apr 5];10(1):1–10. Available from: https://www.nature.com/articles/s41398-020-0800-3

29. Markunas CA, Hancock DB, Xu Z, Quach BC, Fang F, Sandler DP, et al. Epigenome-wide analysis uncovers a blood-based DNA methylation biomarker of lifetime cannabis use. Am J Med Genet B Neuropsychiatr Genet [Internet]. 2021 Apr 1 [cited 2022 May 5];186(3):173–82. Available from: https://pubmed.ncbi.nlm.nih.gov/32803843/

30. Kumar J. Epidemiology of hypertension. Clin Queries Nephrol [Internet]. 2013 Apr 1 [cited 2022 Jul 25];2(2):56–61. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2211947713000162

31. Singh S, Shankar R, Singh GP. Prevalence and Associated Risk Factors of Hypertension: A Cross-Sectional Study in Urban Varanasi. Int J Hypertens [Internet]. 2017 [cited 2022 Jul 25];2017. Available from: https://pubmed.ncbi.nlm.nih.gov/29348933/

32. Minzer S, Losno RA, Casas R. The Effect of Alcohol on Cardiovascular Risk Factors: Is There New Information? Nutrients [Internet]. 2020 Apr 1 [cited 2022 Jun 20];12(4):1–22. Available from: /pmc/articles/PMC7230699/

33. Tasnim S, Tang C, Musini VM, Wright JM. Effect of alcohol on blood pressure. Cochrane Database Syst Rev [Internet]. 2020 Jul 1 [cited 2022 Jul 18];2020(7). Available from: /pmc/articles/PMC8130994/

34. Virdis A, Giannarelli C, Fritsch Neves M, Taddei S, Ghiadoni L. Cigarette smoking and hypertension. Curr Pharm Des [Internet]. 2010 Aug 6 [cited 2022 Jun 20];16(23):2518–25. Available from: https://pubmed.ncbi.nlm.nih.gov/20550499/

35. Dikalov S, Itani H, Richmond B, Vergeade A, Jamshedur Rahman SM, Boutaud O, et al. Tobacco smoking induces cardiovascular mitochondrial oxidative stress, promotes endothelial dysfunction, and enhances hypertension. Am J Physiol Heart Circ Physiol [Internet]. 2019 Mar 1 [cited 2022 Jun 20];316(3):H639–46. Available from: https://pubmed.ncbi.nlm.nih.gov/30608177/

36. Abuhasira R, Haviv YS, Leiba M, Leiba A, Ryvo L, Novack V. Cannabis is associated with blood pressure reduction in older adults - A 24-hours ambulatory blood pressure monitoring study. Eur J Intern Med [Internet]. 2021 Apr 1 [cited 2022 Jul 25];86:79–85. Available from: https://pubmed.ncbi.nlm.nih.gov/33483174/

37. Alshaarawy O, Elbaz HA. Cannabis use and blood pressure levels: United States National Health and Nutrition Examination Survey, 2005-2012. J Hypertens [Internet]. 2016 Aug 1 [cited 2022 Jul 25];34(8):1507–12. Available from: https://pubmed.ncbi.nlm.nih.gov/27270185/

38. Jones RT. Cardiovascular system effects of marijuana. J Clin Pharmacol [Internet]. 2002 Nov 1 [cited 2022 Jul 25];42(S1). Available from: https://pubmed.ncbi.nlm.nih.gov/12412837/

39. Min JL, Hemani G, Smith GD, Relton C, Suderman M. Meffil: efficient normalization and analysis of very large DNA methylation datasets. Bioinformatics [Internet]. 2018 Dec 1 [cited 2022 Jun 23];34(23):3983–9. Available from: https://pubmed.ncbi.nlm.nih.gov/29931280/

40. Zhou W. Infinium Annotation. https://zwdzwd.github.io/InfiniumAnnotation. Accessed 14 June 2022.

41. Chen J, Behnam E, Huang J, Moffatt MF, Schaid DJ, Liang L, et al. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. BMC Genomics [Internet]. 2017 May 26 [cited 2021 Mar 9];18(1):413. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3808-1

42. Ruiz-Arenas C, Gonzalez JR. MEAL: Perform methylation analysis. R package version 1.22.0. 2019 [cited 2020 Apr 9]; Available from: https://bioconductor.org/packages/release/bioc/html/MEAL.html

43. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for {RNA}-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):[e47].

44. Guintivano J, Shabalin AA, Chan RF, Rubinow DR, Sullivan PF, Meltzer-Brody S, et al. Test-statistic inflation in methylome-wide association studies. Epigenetics [Internet]. 2020 Nov 1 [cited 2022 Jul 15];15(11):1163–6. Available from: https://www.tandfonline.com/doi/abs/10.1080/15592294.2020.1758382

45. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes [Internet]. Vol. 27, Nucleic Acids Research. 1999 [cited 2020 Apr 9]. Available from: https://academic.oup.com/nar/article-abstract/27/1/29/1238108

46. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. Vol. 25, Nature Genetics. NIH Public Access; 2000. p. 25–9.

47. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omi A J Integr Biol [Internet]. 2012;16(5):284–7. Available from: https://doi.org/10.1089/omi.2011.0118

48. Piñero J, Piñero P, Bravo A, Uria Queralt-Rosinach N´, Gutí Errez-Sacristán A, Sacristán S, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res [Internet]. 2016 [cited 2020 Apr 9];45:833–9. Available from: http://www.disgenet.org

49. Boca SM, Heller R, Sampson JN. MultiMed: Testing multiple biological mediators simultaneously. 2019.

50. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. mediation: R Package for Causal Mediation Analysis. J Stat Softw [Internet]. 2014;59(5):1–38. Available from: http://www.jstatsoft.org/v59/i05/

51. Battram T, Yousefi P, Crawford G, Prince C, Sheikhali Babaei M, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. Wellcome open Res [Internet]. 2022 May 31 [cited 2022 Jul 4];7:41. Available from: https://pubmed.ncbi.nlm.nih.gov/35592546/

52. Montojo J, Zuberi K, Rodriguez H, Bader GD, Morris Q. GeneMANIA: Fast gene network construction and function prediction for Cytoscape. F1000Research [Internet]. 2014 Jul 1 [cited 2022 Jul 11];3. Available from: /pmc/articles/PMC4168749/

53. Dickson EJ. Recent advances in understanding phosphoinositide signaling in the nervous system. F1000Research [Internet]. 2019 [cited 2022 Jul 15];8. Available from: /pmc/articles/PMC6415330/

54. Enokida H, Shiina H, Urakami S, Terashima M, Ogishima T, Li LC, et al. Smoking influences aberrant CpG hypermethylation of multiple genes in human prostate carcinoma. Cancer [Internet]. 2006 Jan 1 [cited 2022 Jul 8];106(1):79–86. Available from: https://pubmed.ncbi.nlm.nih.gov/16323173/

55. Monick MM, Beach SRH, Plume J, Sears R, Gerrard M, Brody GH, et al. Coordinated Changes in AHRR Methylation in Lymphoblasts and Pulmonary Macrophages from Smokers. Am J Med Genet [Internet]. 2012 [cited 2022 Jul 8];159B(2):141. Available from: /pmc/articles/PMC3318996/

56. Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet [Internet]. 2012 Jul

[cited 2022 Jul 8];21(13):3073–82. Available from: https://pubmed.ncbi.nlm.nih.gov/22492999/

57. Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet [Internet]. 2013 Mar [cited 2022 Jul 8];22(5):843–51. Available from: https://pubmed.ncbi.nlm.nih.gov/23175441/

58. Wider C, Lincoln SJ, Heckman MG, Diehl NN, Stone JT, Haugarvoll K, et al. Phactr2 and Parkinson's disease. Neurosci Lett [Internet]. 2009 Mar 3 [cited 2022 Jul 26];453(1):9. Available from: /pmc/articles/PMC2684848/

59. Liang X, Justice AC, So-Armah K, Krystal JH, Sinha R, Xu K. DNA methylation signature on phosphatidylethanol, not on self-reported alcohol consumption, predicts hazardous alcohol consumption in two distinct populations. Mol Psychiatry [Internet]. 2021 Jun 1 [cited 2022 May 12];26(6):2238–53. Available from: https://pubmed.ncbi.nlm.nih.gov/32034291/

60. Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry 2018 232 [Internet]. 2016 Nov 15 [cited 2022 May 13];23(2):422–33. Available from: https://www.nature.com/articles/mp2016192

61. Aragona J, Lee CK. Scoliosis in fetal alcohol syndrome: A case report. Orthopedics. 1981;4(10):1141–3.

62. Gallagher C, McCarthy FP, Ryan RM, Khashan AS. Maternal Alcohol Consumption During Pregnancy and the Risk of Autism Spectrum Disorders in Offspring: A Retrospective Analysis of the Millennium Cohort Study. J Autism Dev Disord [Internet]. 2018 Nov 1 [cited 2022 Jul 26];48(11):3773. Available from: /pmc/articles/PMC6182718/

63. Singer AB, Aylsworth AS, Cordero C, Croen LA, DiGuiseppi C, Fallin MD, et al. Prenatal Alcohol Exposure in Relation to Autism Spectrum Disorder: Findings from the Study to Explore Early Development (SEED). Paediatr Perinat Epidemiol [Internet]. 2017 Nov 1 [cited 2022 Jul 26];31(6):573. Available from: /pmc/articles/PMC5690833/

64. Varadinova M, Boyadjieva N. Epigenetic mechanisms: A possible link between autism spectrum disorders and fetal alcohol spectrum disorders. Pharmacol Res. 2015 Dec 1;102:71–80.

65. Watson CT, Szutorisz H, Garg P, Martin Q, Landry JA, Sharp AJ, et al. Genome-Wide DNA Methylation Profiling Reveals Epigenetic Changes in the Rat Nucleus Accumbens Associated With Cross-Generational Effects of Adolescent THC Exposure. Neuropsychopharmacology [Internet]. 2015 Dec 1 [cited 2022 May 13];40(13):2993–3005. Available from: https://pubmed.ncbi.nlm.nih.gov/26044905/

66. Allen JP, Danoff JS, Costello MA, Hunt GL, Hellwig AF, Krol KM, et al. Lifetime marijuana use and epigenetic age acceleration: A 17-year prospective

examination. Drug Alcohol Depend [Internet]. 2022 Apr [cited 2022 Apr 5];233:109363. Available from: https://pubmed.ncbi.nlm.nih.gov/35231715/

67. Tantoh DM, Lee KJ, Nfor ON, Liaw YC, Lin C, Chu HW, et al. Methylation at cg05575921 of a smoking-related gene (AHRR) in non-smoking Taiwanese adults residing in areas with different PM 2.5 concentrations. Clin Epigenetics [Internet]. 2019 May 6 [cited 2022 Jul 8];11(1). Available from: https://pubmed.ncbi.nlm.nih.gov/31060609/

68. Zhang Y, Zeng C. Role of DNA methylation in cardiovascular diseases. Clin Exp Hypertens [Internet]. 2016 Apr 2 [cited 2022 Jul 21];38(3):261–7. Available from: https://pubmed.ncbi.nlm.nih.gov/27028400/

69. Gonzalez-Jaramillo V, Portilla-Fernandez E, Glisic M, Voortman T, Bramer W, Chowdhury R, et al. The role of DNA methylation and histone modifications in blood pressure: a systematic review. J Hum Hypertens [Internet]. 2019 Oct 1 [cited 2022 Jul 11];33(10):703–15. Available from: https://pubmed.ncbi.nlm.nih.gov/31346255/

70. Richard MA, Huan T, Ligthart S, Gondalia R, Jhun MA, Brody JA, et al. DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. Am J Hum Genet [Internet]. 2017 Dec 7 [cited 2022 Jul 11];101(6):888–902. Available from: https://pubmed.ncbi.nlm.nih.gov/29198723/

71. Syme C, Shin J, Richer L, Gaudet D, Fornage M, Paus T, et al. Epigenetic Loci of Blood Pressure. Circ Genomic Precis Med [Internet]. 2019 Jan 1 [cited 2022 Jul 11];12(1):E002341. Available from: https://pubmed.ncbi.nlm.nih.gov/30645168/

5.2. Manuscript 2

The early-life exposome modulates the effect of polymorphic inversions on DNA methylation

Carreras-Gallo N [1], Cáceres A [1], Balagué-Dobón L, Ruiz-Arenas C, Andrusaityte S, Carracedo A, Casas M, Chatzi L, Grazuleviciene R, Gutzkow KB, Lepeule J, Maitre L, Nieuwenhuijsen M, Slama R, Stratakis N, Thomsen C, Urquiza J, Wright J, Yang T, Escaramís G, Bustamante M, Vrijheid M, Pérez-Jurado LA, González JR

[1] These authors contributed equally.

The early-life exposome modulates the effect of polymorphic inversions on DNA methylation. *Communications Biology*. 5, 455 (2022). https://doi.org/10.1038/s42003-022-03380-2. IF: 6.5. Position: Q1 - D1.

Supplementary Material here

communications biology

ARTICLE

https://doi.org/10.1038/s42003-022-03380-2 OPEN

The early-life exposome modulates the effect of polymorphic inversions on DNA methylation

Natàlia Carreras-Gallo[1,†], Alejandro Cáceres[1,2,3,†], Laura Balagué-Dobón[1], Carlos Ruiz-Arenas[4,5,6], Sandra Andrusaityte[7], Ángel Carracedo[8,9], Maribel Casas[1,2,6], Leda Chatzi[10], Regina Grazuleviciene[7], Kristine Bjerve Gutzkow[11], Johanna Lepeule[12], Léa Maitre[1,2,6], Mark Nieuwenhuijsen[1,2,6], Remy Slama[12], Nikos Stratakis[1], Cathrine Thomsen[11], José Urquiza[1,2,6], John Wright[13], Tiffany Yang[13], Geòrgia Escaramís[2,14,15], Mariona Bustamante[1,2,6,16], Martine Vrijheid[1,2,6], Luis A Pérez-Jurado[4,5,6,17], Juan R González[1,2,18,*]

Polymorphic genomic inversions are chromosomal variants with intrinsic variability that play important roles in evolution, environmental adaptation, and complex traits. We investigated the DNA methylation patterns of three common human inversions, at 8p23.1, 16p11.2, and 17q21.31 in 1,009 blood samples from children from the Human Early Life Exposome (HELIX) project and in 39 prenatal heart tissue samples. We found inversion-state specific methylation patterns within and nearby flanking each inversion region in both datasets. Additionally, numerous inversion-exposure interactions on methylation levels were identified from early-life exposome data comprising 64 exposures. For instance, children homozygous at inv-8p23.1 and higher meat intake were more susceptible to *TDH* hypermethylation (P=3.8x10^{-22}); being the inversion, exposure, and gene known risk factors for adult obesity. Inv-8p23.1 associated hypermethylation of *GATA4* was also detected across numerous exposures. Our data suggests that the pleiotropic influence of inversions during development and lifetime could be substantially mediated by allele-specific methylation patterns which can be modulated by the exposome.

[1] Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain. [2] Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. [3] Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya, Barcelona 08019, Spain. [4] Institut Hospital del Mar d'Investigacions Mediques (IMIM), Barcelona, Spain. [5] Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain. [6] Department of Health and Experimental Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain. [7] Department of Environmental Science, Vytautas Magnus University, 44248 Kaunas, Lithuania. [8] Medicine Genomics Group, Centro de Investigación Biomédica en Red Enfermedades Raras (CIBERER), University of Santiago de Compostela, CEGEN-PRB3, Santiago de Compostela, Spain. [9] Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Servicio Gallego de Salud (SERGAS), Santiago de Compostela, Galicia. [10] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA. [11] Department of Environmental Health, Norwegian Institute of Public Health, 0456, Oslo, Norway. [12] Institut national de la santé et de la recherche médicale (Inserm) and Université Grenoble-Alpes, Institute for Advanced Biosciences (IAB), Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Grenoble, France. [13] Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK. [14] Department of Biomedical Science, Faculty of Medicine and Health Science, University of Barcelona, Barcelona, Spain. [15] Research Group on Statistics, Econometrics and Health (GRECS), UdG, Girona, Spain. [16] Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. [17] Genetics Service, Hospital del Mar, Barcelona, Spain. [18] Department of Mathematics, Universitat Autònoma de Barcelona, Bellaterra, Spain. † These authors contributed equally. * Email: juranr.gonzalez@isglobal.org

Inversions are segments of DNA that run in the opposite direction to a reference genome. They are balanced mutations of different sizes, from a gene's exon to a chromosome's portion[1]. Because of their role in adaptation to the environment, chromosome evolution, and sex determination systems in multiple species, polymorphic inversions have traditionally displayed a great interest in evolutionary biology[2,3]. Recent studies have shown that they are important contributors to the genetic basis of common complex diseases in humans, such as obesity, diabetes, asthma, cancer, and neurological conditions such as depression or neuroticism[4–11]. By capturing multiple functional variants, inversions can confer simultaneous risks to different diseases, and, as such, increase the frequency of the diseases' co-morbidities. Human inversions at 8p23.1, 16p11.2, and 17q21.31 are large, common, and associate with multiple diseases, including those co-occurring with obesity[5,8]. In addition, they have been strongly correlated with the expression of the several genes they encapsulate across multiple tissues[8,12–14]. There are different mechanisms from which inversions can modulate gene expression. First, inversions can break genes or displace regulatory elements with important functional and phenotypic consequences[10,12,15]. Second, recombination is suppressed in the inverted region in heterokaryotypes. As such, inverted and non-inverted alleles accumulate different genetic variants that support differences of gene expression between alleles[2,16,17]. Although several studies have demonstrated the effect of inversions on gene expression, it is unknown the extent to which inversions are also characterized by specific methylation patterns.

DNA methylation, the addition of a methyl group in a CpG DNA site, plays an important and complex role in the regulation of gene expression[18]. Depending on the relative position of the CpG site within the gene, its methylation can increase or decrease the gene's expression[19]. Methylated

promoters are often associated with deactivation of transcription, while methylation within the gene's body avoids alternative start sites[20]. Methylation is often strongly correlated across contiguous CpG sites, a fact that is used to determine differentially methylated regions (DMR) of kilobase pair lengths[21]. At larger distances, coherent methylation patterns may be supported by genomic variants such as copy number variants[22]. However, it is unknown if methylation patterns in inverted regions can also be detected. We, therefore, hypothesized that the common human inversions at 8p23.1, 16p11.2, and 17q21.31 are correlated with the methylation of multiple CpG sites within and surrounding the inverted region, creating allele-specific methylation patterns. In support of this hypothesis, some studies have already reported associations between inversion and phenotypes likely modulated by specific methylation changes[6,23,24]. Besides, since CpG methylation is involved in regulating chromatin structure[25], these methylation patterns could be associated with different tridimensional (3D) DNA structures for each allele. This would be in line with the influence on 3D DNA structure by large structural variants reported by Shanta et. al[26].

Results

Frequency of inversions at 8p23.1, 16p11.2, and 17q21.31

We analyzed data from the Human Early Life Exposome (HELIX) project, a multicenter European cohort (Spain, UK, France, Lithuania, Norway, and Greece). This project comprises 1,301 children with genomic, transcriptomic, epigenomic, and exposome data[36]. HELIX has the goal of characterizing the exposome during early-life and evaluating its relationship with molecular signatures and child health outcomes. The genome-wide blood DNA methylation and blood cell transcriptome were measured at the ages between 6 and 11. From this dataset, we selected

children with genetic and methylation data. We used *Peddy*[37] to estimate major population ancestry groups and individuals of European ancestry were kept in the analysis, resulting in a total of 1,009 children included in the analyses.

We called 8p23.1, 16p11.2, and 17q21.31 inversion genotypes from the selected children using *scoreInvHap*[11] on imputed SNP array data. Inversion genotypes were labeled as N/N for non-inverted homozygous; N/I for heterozygous; and I/I for inverted homozygous. We observed that the frequencies for the inverted allele were consistent with those reported for Europeans (55.70%, 35.70%, and 21.95% for inversions at 8p23.1, 16p11.2, and 17q21.31, respectively)[1,11]. As expected, we did not observe significant variation between sexes (Supplementary Figure 1a-c), but we observed some variations across cohorts (Supplementary Figure 1d-f). As previously reported[8], we evaluated the south-north gradient for the inverted allele frequency and we observed a positive correlation for inv-16p11.2 ($r = 0.79$, $P = 0.058$), and a negative correlation for inv-17q21.3 ($r = - 0.92$, $P = 0.009$) (Supplementary Figure 2). For the inv-8p23.1, we did not observe a significant south-north gradient ($r = - 0.33$, $P = 0.519$).

Inversions as *eQTL*s in blood cells

We first evaluated the inversion status as expression quantitative loci (*eQTL*) of the genes within the inversion regions +/- 1 Mb. We performed the association analyses of the inversions in each separate cohort adjusting by sex, age, cell-type proportions (inferred from methylation data), and 10 genome-wide principal components of genomic SNP variation (N=790). We then combined the results with a meta-analysis across cohorts. Results were considered significant when they passed Bonferroni's correction for multiple comparisons. We confirmed that the inv-8p23,1 and inv-16p11.2 were *eQTL*s for the numerous neighboring genes and the genes they

encapsulate (see Supplementary Data 1 and Supplementary Figure 3). We observed 12 genes that were significantly associated with inv-8p23.1. We detected significant up-regulation of *BLK, SLC35G5/SLC35G4, FAM86B1/FAM86B2,* and *FAM86B3P*, and down-regulation of *FDFT1, FAM167A, FAM66D, SGK223, XKR6*, and *LOC100506990* for the inverted allele. In the case of the polymorphic inversion at 16p11.2, we observed 10 significant associations including up-regulation of *TUFM, MIR4721, EIF3C/EIF3CL, LAT, SPNS1,* and *NPIPB9/NPIPB8/NPIPB7* for the inverted allele and down-regulation of *SGF29, SBK1, LOC388242,* and *SULT1A1*. Finally, for inv-17q21.31, we did not observe *eQTL* effects, perhaps because single-copy genes within this inversion are mostly expressed in the brain[14]. We thus confirmed the effect of the inversions 8p23.1 and 16p11.2 on the gene expression in blood in 6-11 year-old children, as previously observed in adults across different tissues[8,12–14].

Inversions as *mQTL*s in blood cells

We then studied the associations of the genotypes of each of the three inversions with the differential methylation of CpG sites within the +/- 1 Mb regions containing the inversions (Supplementary Data 2). We removed CpG sites with single nucleotide polymorphic (SNP) variation. We performed the analyses in each separate cohort adjusting by the same covariates likewise the transcription analyses. We combined the results with a meta-analysis across cohorts (N=1,009). As illustrated in Figure 1a-c, all three inversions were significantly associated with differences in methylation across multiple CpG sites after Bonferroni's correction for multiple comparisons. We also observed that the most significant associations were in CpG sites within the inversion region or close to the breakpoints. In particular, we observed that 15.21% (129 of 848) CpG sites within and around inv-8p23.1 had significant differences in methylation

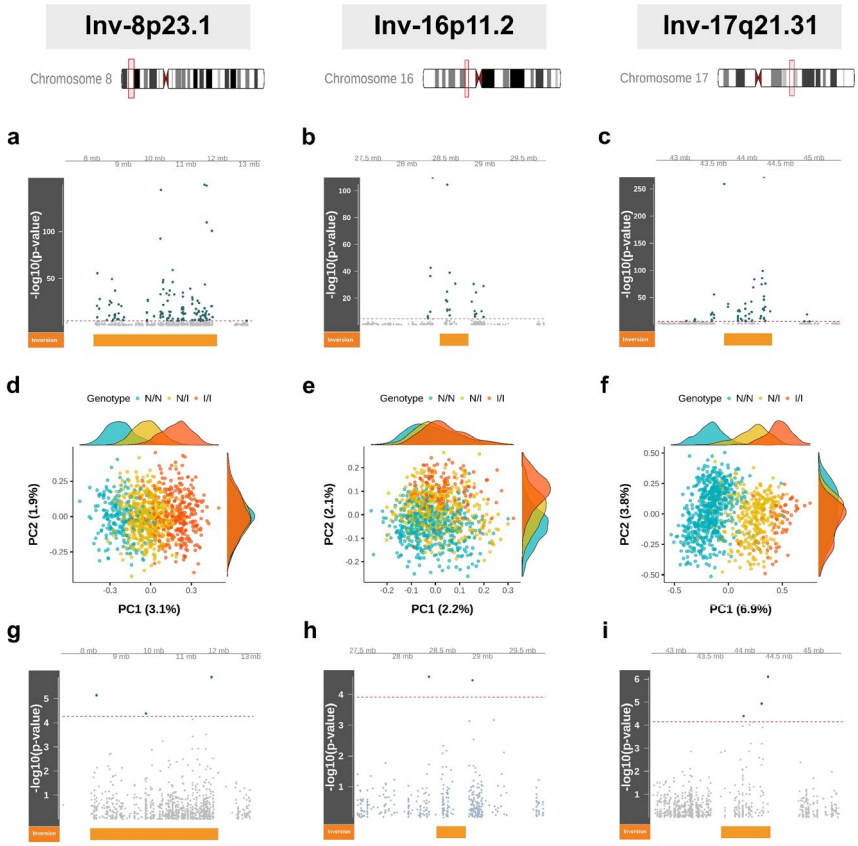


Figure 1 | Inversion status as methylation quantitative trait locus (*mQTL*) of multiple CpG sites within and surrounding three common human inversions. The first column in the plot panel corresponds to inv-8p23.1, the second to inv-16p11.2, and the third to inv-17q21.31. **a-c**) Manhattan plots for the significance of the associations between the differential methylation of the CpG sites and the inversion genotypes in child blood cells (N=1,009). The x-axes show the chromosome position (±1Mb between the inversions' breakpoints). The y-axes show the $-\log_{10}$(*P*-value). The dashed red line indicates Bonferroni's threshold of significance. Green points are CpG sites with significant associations and those in grey are non-significant. The orange block illustrates the inversions' region. **d-f**) Principal Component (PC) analysis for methylation levels of CpG sites within and surrounding the inversions, revealing remarkably distinctive methylation patterns among the different inversion statuses. Blue points illustrate non-inverted homozygous (N/N), yellow illustrate heterozygous (N/I), and orange illustrate inverted homozygous (I/I) individuals. In parenthesis, the methylation variance explained by each PC. **g-i**) Manhattan plots of differentially methylated CpG sites depending on the inversion genotypes in fetal heart DNA (N=40).

levels according to the inversion status (min. $P = 63.1 \times 10^{-147}$, Figure 1a), with 49 significant CpG sites hypermethylated and 80 hypomethylated in the inverted concerting the non-inverted allele. For this inversion, we observed 24 genes with at least one significant differentially methylated CpG site and five genes with more than five differentially methylated sites; namely *MSRA, MFHAS1, BLK, RP1L1,* and *XKR6*. For inv-16p11.2, we found 27 significant CpG sites differentially methylated from a total of 401 (6.73%, min. $P < 10^{-300}$, Figure 1b), with 9 significant CpG sites hypermethylated and 18 hypomethylated at the inverted allele. For this inversion, we observed 11 genes with at least one significant CpG site. *IL27* was the gene with the greatest number of CpG sites (5) differentially methylated (all hypomethylated at the inverted allele). Finally, 58 CpG sites from 666 (8.71%, min. $P < 10^{-300}$, Figure 1c) had significant methylation differences for inv-17q21.31 (30 hypermethylated and 28 hypomethylated at the inverted allele). *CRHR1, MAPT*, and *KANSL1* were the 17q21.31 genes with the highest number of differentially methylated CpG sites and a total of 14 genes had at least one CpG site differentially methylated. Therefore, each of these three inversions behaves as an extended methylation quantitative trait locus (*mQTL*) covering hundreds of kilo-bases, an observation that had not been previously reported.

To establish the degree to which the association between the effect of inversion status on CpG methylation is associated with changes in gene expression of surrounding genes, we searched for the methylation changes that locate in differentially expressed genes (Supplementary Figure 4). We observed that four genes (*BLK, FDFT1, XKR6*, and *FAM167A*) overlapped for the inv-8p23.1 with differentially methylated CpG sites. We analyzed whether the observed expression changes were in the expected directions based on the methylation of these regions; that is, hypermethylation of the promoters for downregulated genes, hypomethylation of the promoters for

upregulated genes, and hypermethylation of the bodies for upregulated genes. *XKR6* was a highly consistent case whose downregulation and methylation, across 11 CpG sites within its body, were associated with the inverted allele. For inv-16p11.2, we observed four genes that were differentially expressed and methylated by the inversion allele (*TUFM, SBK1, SPNS1*, and *SULT1A1*). In this case, most of the CpG sites were in the promoter region (TSS1500) and the relation between the expression and methylation levels was consistent. We further observed that *SULT1A1* and *TUFM* had CpG sites in their promoters (cg01378222 and cg00348858) that highly associated with the effect of inversion in gene expression. We found that cg01378222 mediated the 95% of the association between inv-16p11.2 and the expression of *SULT1A1* ($P < 2x10^{-16}$), and that cg00348858 mediated the 5% of the association between the inversion and *TUFM* expression ($P = 0.002$).

These findings provided evidence of regulatory pathways where inversion, methylation, and gene expression are all involved. In addition, our observation that inv-17q21.31 did not show *eQTL* effects in blood indicates that the three-way association of the variables is tissue specific, as we observed a clear methylation pattern for the inversion.

Inversion-state specific methylation patterns

In order to define whether the methylation patterns were specific to each inversion allele, we performed principal component (PC) analysis of the methylation levels of CpG sites within and around each inversion. We thus quantified individual differences in methylation profiles across the inverted regions. We included the region +/- 1 Mb to account for the effect of the inversions beyond the breakpoints. Remarkably, the first component strongly correlated with the inversion genotype of the individuals in all three inversions (inv-8p23.1 PC 1: R2 = 0.68, $P < 2x10^{-16}$, inv-16p11.2 PC

1: R2 = 0.05, $P = 1.34x10^{-12}$, and inv-17q21.31 PC 1: R2 = 0.70, $P < 2x10^{-16}$), see Figure 1d-f. We observed that the first PC clearly separated the genotypes of inversions at 8p23.1 and 17q21.31, possibly sustained by the haplotypic differences between inversion status. While the first PC of inv-16p11.2 was significantly associated with inversion genotypes, the second PC was also needed to distinctly separate the genotypes (R2 = 0.33, $P < 2x10^{-16}$). This is in line with the univariate differential analysis, where inv-16p23.1 showed the smallest proportion of CpG sites differentially methylated according to the inversion status. This is possibly explained by the multiple haplotypes supported by this inversion[11]. These analyses showed that hyper and hypomethylation patterns of CpG sites across the inverted regions are specific to the inversion status.

Inversions as *mQTL*s in fetal heart DNA

We asked whether the effect of the inversion on DNA methylation could be also seen prenatally and in another tissue. Using methylation data of heart DNA from 39 fetuses from interrupted pregnancies at 21-22 weeks of gestational age due to congenital heart defects[38], we performed the same differential analysis adjusting by sex. We observed that all the inversions act as *mQTL*s during early development from conception, although few CpG sites per inversion passed Bonferroni's threshold (Figure 1g-i and Supplementary Data 3). This can be explained by the small sample size. Nonetheless, we observed that the distribution of the significant associations was very similar to the one observed in HELIX data, having greater differences in methylation in the CpG sites between the breakpoints. Additionally, we saw that 38 CpG significant sites overlapped between heart (nominal *P*-value) and blood (adjusted *P*-value) tissues, 32 of which were in the same direction, suggesting that the effect of inversions on CpG methylation may be sustained between tissues and stages of life.

Effect of inversion-exposure interactions on DNA methylation

As these common human inversions at 8p23.1, 16p11.2, and 17q21.31 offered a solid genetic context where allele-specific methylation patterns were found, we then asked whether these patterns were modulated by environmental exposures. Thus, we assessed which of 64 exposures at early-life differentially modified the methylation levels of the CpG sites within the inversion regions according to the inversion status.

We performed differential methylation analyses for the interactions of the 3 inversions with 64 exposures (7 during pregnancy and 57 at 6-11 years of age) grouped by 12 exposure families, including build environment, air pollution, persistent and non-persistent chemicals, diet, and exposure to tobacco smoke, among others (Figure 2a and Supplementary Data 4). We observed 36 exposures and 58 CpG sites implicated in at least one significant inversion-exposure interaction after Bonferroni's correction for multiple comparisons (see Table 1 and Supplementary Data 5). All exposure families had at least one exposure that interacted with one of the three inversions, except natural spaces and polybrominated diphenyl ether compounds (PBDE). Remarkably, the exposure families with the greatest number of significant interactions were metals (13 interactions), diet (11), phenols (11), and organochlorines (OCs) (10) (Supplementary Data 6).

Inversion at 8p23.1 had 36 significant interactions with exposures from 9 different families (Figure 2b). OC was the most predominant exposure family involved in 8 interactions, followed by diet with 6 and phenols with 5. The genes with the greatest number of CpG sites differentially methylated according to the interactions were *GATA4* (hypomethylated for the inverted allele in all but one), *XKR6* (hypermethylated for the inverted allele in all but one), *TDH*, and *FAM167A*, all of them seen differentially methylated depending on the inversion haplotype. In the case of inv-

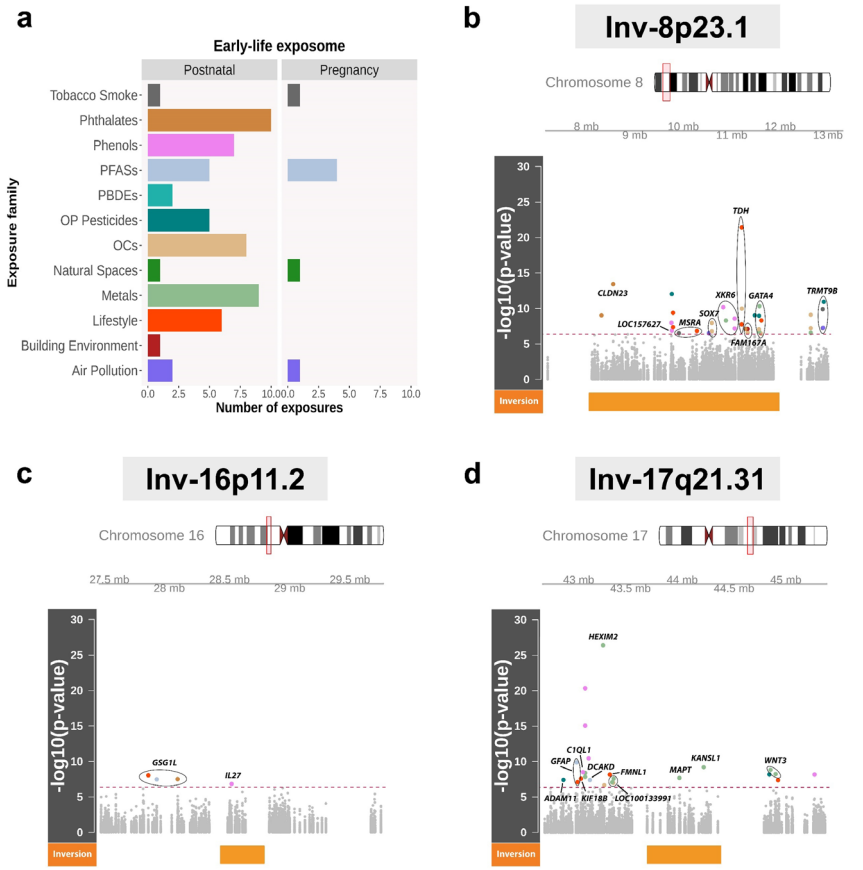


Figure 2 | Inversion-exposure interactions as methylation quantitative trait locus (*mQTL*) of multiple CpG sites within and surrounding three common human inversions. a) Number of exposures per family in the early-life exposome from the HELIX project. **b-d**) Manhattan plots showing the significance of the associations (N=1,009) between the differential methylation of the CpG sites and the inversion-exposure interactions across all 64 exposures (**a**) and the genotypes of three human inversions at 8p.23.1 (**b**); 16p11.2 (**c**) and 17q21.31 (**d**), illustrated by the orange block. The x-axes show the chromosome position (±1Mb between the inversions' breakpoints). The y-axes show the $-\log_{10}$ (*P*-value) of the associations. The dashed red line indicates Bonferroni's threshold of significance. Significant results are colored according to the family exposure (**a**) and labeled according to the closest gene to the CpG (Illumina annotation). Grey points are not significant.

Table 1 | Significant associations between CpG methylation levels and inversion-exposure interactions.

Exposure	Exposure Family	Period	Inversion	CpG	Location	Gene Symbol	Effect	*P*-value
Lead	Metals	Postnatal	17q21.31	cg19655070	chr17:43237981	*HEXIM2*	-0.043	4.5E-27
Meat intake	Diet	Postnatal	8p23.1	cg01489256	chr8:11204017	*TDH*	0.0156	3.8E-22
MEPA	Phenols	Postnatal	17q21.31	cg06368300	chr17:43065840		0.0077	5.1E-21
MEPA	Phenols	Postnatal	17q21.31	cg11178337	chr17:43065745		0.0189	9E-16
MBzP	Phthalates	Postnatal	8p23.1	cg06671706	chr8:8559999	*CLDN23*	0.0173	3.8E-14
DETP	OP Pesticides	Postnatal	8p23.1	cg17526103	chr8:9765691		0.0038	9.5E-13
DMTP	OP Pesticides	Postnatal	8p23.1	cg17120402	chr8:12891262		0.0065	1.2E-11
MEPA	Phenols	Postnatal	17q21.31	cg07822074	chr17:43098904		0.0049	3.6E-11
Manganese	Metals	Postnatal	8p23.1	cg26020513	chr8:11568356	*GATA4*	-0.033	4.8E-11
OXBE	Phenols	Postnatal	8p23.1	cg20858107	chr8:10823238	*XKR6*	-0.004	6.7E-11
HCB	OCs	Postnatal	8p23.1	cg03399933	chr8:11205972	*TDH*	-0.023	1.1E-10
Parental smoking	Tobacco Smoke	Postnatal	8p23.1	cg08196601	chr8:12869553	*TRMT9B*	-0.01	1.3E-10
PFUNDA	PFASs	Postnatal	17q21.31	cg23016243	chr17:42983768	*GFAP*	-0.004	1.3E-10
KIDMED score	Diet	Postnatal	8p23.1	cg19352062	chr8:9791449		0.0054	4.1E-10
Molybdenum	Metals	Postnatal	17q21.31	cg13465858	chr17:44204908	*KANSL1*	0.0217	6.3E-10
PCB 180	OCs	Postnatal	8p23.1	cg19931644	chr8:12623485		0.0185	7.9E-10
DMDTP	OP Pesticides	Postnatal	8p23.1	cg07291889	chr8:11471712		-0.014	9.6E-10
MBzP	Phthalates	Postnatal	8p23.1	cg19996406	chr8:8318774		-0.008	9.7E-10
DEP	OP Pesticides	Postnatal	8p23.1	cg22320962	chr8:11560299	*GATA4*	-0.005	1.1E-09
Molybdenum	Metals	Postnatal	17q21.31	cg16677019	chr17:44847268	*WNT3*	-0.02	1.5E-09
ETPA	Phenols	Postnatal	8p23.1	cg11051055	chr8:11058145	*XKR6*	0.0076	2.8E-09
ETPA	Phenols	Postnatal	17q21.31	cg24945657	chr17:43044484	*C1QL1*	-0.011	3.2E-09
Arsenic	Metals	Postnatal	17q21.31	cg06368300	chr17:43065840		0.0077	4.1E-09
KIDMED score	Diet	Postnatal	8p23.1	cg12395012	chr8:11607386	*GATA4*	-0.004	5.1E-09
Cadmium	Metals	Postnatal	8p23.1	cg02569740	chr8:10878898	*XKR6*	0.0093	5.2E-09
Mercury	Metals	Postnatal	17q21.31	cg16440629	chr17:44896147	*WNT3*	0.0073	6E-09
DEP	OP Pesticides	Postnatal	17q21.31	cg23968286	chr17:44835681		-0.004	6.7E-09
OXBE	Phenols	Postnatal	17q21.31	cg07673979	chr17:45270216		-0.003	6.9E-09
KIDMED score	Diet	Postnatal	17q21.31	cg09264140	chr17:43302776	*FMNL1*	-0.005	7E-09
Vegetables intake	Diet	Postnatal	16p11.2	cg08755784	chr16:27829728	*GSG1L*	0.0065	8.9E-09
ETPA	Phenols	Postnatal	8p23.1	cg01454752	chr8:9758847	*LOC157627*	0.0078	1.1E-08
HCB	OCs	Postnatal	8p23.1	cg24690731	chr8:10589093	*SOX7*	-0.02	1.1E-08
Cobalt	Metals	Postnatal	17q21.31	cg06368300	chr17:43065840		-0.022	1.4E-08

Table 1 (continuation) | Significant associations between CpG methylation levels and inversion-exposure interactions.

Exposure	Exposure Family	Period	Inversion	CpG	Location	Gene Symbol	Effect	P-value
Meat intake	Diet	Postnatal	8p23.1	cg02601489	chr8:11203954	*TDH*	0.0092	1.8E-08
Copper	Metals	Postnatal	17q21.31	cg05301556	chr17:43971177	*MAPT; LOC100128977*	0.0522	2E-08
Cobalt	Metals	Postnatal	17q21.31	cg26742995	chr17:43339594	*LOC100133991; SPATA32*	0.0198	2.6E-08
KIDMED score	Diet	Postnatal	17q21.31	cg00240569	chr17:43025343	*KIF18B*	0.0052	2.6E-08
MEHP	Phthalates	Postnatal	16p11.2	cg03962082	chr16:28072873	*GSG1L*	-0.01	3E-08
PFHXS	PFASs	Pregnancy	16p11.2	cg01896119	chr16:27899404	*GSG1L*	-0.014	3.3E-08
DMTP	OP Pesticides	Postnatal	17q21.31	cg11640208	chr17:42857157	*ADAM11*	-0.006	3.8E-08
PFUNDA	PFASs	Postnatal	17q21.31	cg18176312	chr17:43111632	*DCAKD*	-0.006	4E-08
Fish and seafood intake	Diet	Postnatal	17q21.31	cg17101843	chr17:44919554		-0.01	4.1E-08
Vegetables intake	Diet	Postnatal	8p23.1	cg00056202	chr8:9791350		0.0085	4.4E-08
PM2.5 (preg)	Air Pollution	Pregnancy	8p23.1	cg26339990	chr8:12878608	*TRMT9B*	-0.003	5.5E-08
Active smoking (preg)	Tobacco Smoke	Pregnancy	8p23.1	cg08196601	chr8:12869553	*TRMT9B*	-0.02	5.9E-08

The table illustrates the top 45 significant associations of CpG sites (+/-1Mb) and the interactions of three common human inversions (inv-8p23.1, inv-16p11.2 and inv-17q21.31) with exposures in the HELIX exposomic data. The full table is available in **Supplementary Data 5**. The first column indicates the exposure involved in the interaction (the description of the exposures is detailed in **Supplementary Data 4**). Exposures are described by their families and the period which they were measured. The inversion column describes the inversion interacting with the exposure. CpG sites are described by their name, location, and gene symbol (written in italics), when mapped to a gene. The Effect column represents the estimate of the interaction effect and the P-value column its nominal level of significance.

16p11.2, we only found 4 significant interactions (Figure 2c). Notably, 3 interactions contributed to *GSG1L* methylation changes: child vegetables intake (cg08755784, β = 0.006, $P = 8.9x10^{-9}$), child Mono-2-ethylhexyl phthalate (MEHP) levels (cg03962082; β = -0.011, $P = 3.0x10^{-8}$), and child Perfluorohexane sulfonate (PFHXS) levels (cg01896119; β = -0.014, $P = 3.3x10^{-8}$). For inv-17q21.31, we observed 24 significant interactions with exposures from 6 exposure families (Figure 2d). The most frequent family was metals with 9 significant interactions with inv-17q21.31. The most significant interaction of the inversion was with the exposure to lead on *HEXIM2* methylation (cg19655070: β = -0.043, $P = 4.5x10^{-27}$). Furthermore, several CpG sites in the up-stream region of *C1QL1* were differentially methylated according to the interaction of inv-17q21.31 with phenols. In particular, a CpG site within *C1QL1* promoter was hypomethylated for the inverted allele when the ethyl paraben (ETPA) exposure increased (cg24945657: β = -0.011, $P = 3.2x10^{-9}$). In addition, three intergenic CpG sites near this gene promoter were hypermethylated for the inverted allele when the exposure to methyl paraben (MEPA) increased (cg06368300: β = 0.008, $P = 5.1x10^{-21}$; cg11178337: β = 0.019, $P = 9.0x10^{-16}$; cg07822074: β = 0.005, $P = 3.6x10^{-11}$). It should be noted that there are four genes (*KANSL1, MAT, LOC100128977,* and *WNT3*) in this region with significant associations that were also differentially methylated depending on the inversion haplotype.

Genes with strongest and most numerous inversion-exposure interactions

Within the significant interactions (Table 1), we looked in detail at the genes that showed both the highest significant levels and multiple interactions across different CpG sites for the same gene. We identified three relevant genes within inv-8p23.1, namely *TDH, GATA4*, and

TRMT9B. Within *TDH* we found two CpG sites significantly associated with the interaction between the inversion and meat intake: cg01489256 ($\beta = 0.0156$, $P = 3.8x10^{-22}$) and cg02601489 ($\beta = 0.0092$, $P = 1.8x10^{-8}$). More specifically, we observed that individuals homozygous for the non-inverted allele (N/N) had a negative association, while heterozygous individuals did not present any association, and homozygous for the inverted allele (I/I) had a positive association (Figure 3a). We also observed that the association was consistent across all the cohorts, with no significant heterogeneity ($P = 0.39$ and $P = 0.45$), see Figure 3b. We further observed that the increase of meat intake reduced the expression of *TDH* ($P = 0.00398$) while the associated methylation effect on the expression depended on the genetic context given by the inversion, adjusting by sex, age, and cohort (CpG-inversion interaction, $P = 0.00193$) (Supplementary Figure 5). Remarkably, the gene, the inversion, and the exposure have been independently associated with obesity in adults[5,39–41].

GATA4 was the gene with the greatest number of CpG sites that changed their methylation according to different interactions between inv-8p23.1 and exposures from different families. These interactions included manganese (cg26020513: $\beta = -0.033$, $P = 4.8x10^{-11}$), diethyl-phosphate (DEP) (cg22320962: $\beta = -0.005$, $P = 1.1x10^{-9}$), Mediterranean Diet Quality Index for children and teenagers (KIDMED) (cg12395012: $\beta = -0.004$, $P = 5.1x10^{-9}$), mercury (cg27100236: $\beta = -0.007$, $P = 1.8x10^{-7}$), and PCB 138 (cg13293535: $\beta = 0.013$, $P = 3.5x10^{-7}$) exposures. We observed that this CpG was hypermethylated in the individuals homozygous for non-inverted allele when increasing the exposure to manganese (Figure 3c). The meta-analysis also revealed consistency across cohorts with no significant heterogeneity ($P = 0.74$) (Figure 3d). Interestingly, hypermethylation of *GATA4* in developing heart DNA, particularly at cg26020513, has been previously associated with congenital heart defects in fetuses[42].

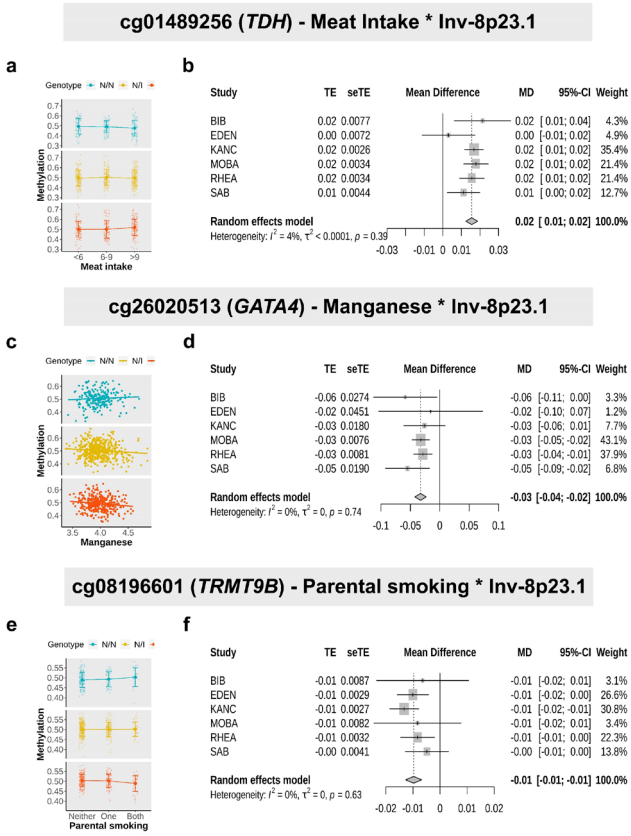


Figure 3 | Interaction and forest plots for *TDH, GATA4,* and *TRMT9B* genes. **a**) Interaction plot illustrating differences across inv-8p23.1 genotypes in the association between cg01489256 (*TDH*) methylation and meat intake (expressed in servings per week). Methylation means given meat intake status and inversion genotype are represented with their 95% confidence intervals (N=1,009). **b**) Forest plot showing the meta-analysis effect estimates of inv-8p23.1-meat intake interaction on cg01489256 methylation across HELIX cohorts. **c**) Interaction plot illustrating differences across inv-8p23.1 genotypes in the association between cg26020513 (*GATA4*) methylation and manganese (N=1,009). **d**) Forest plot showing the meta-analysis effect estimates of inv-8p23.1-manganese interaction on cg26020513 methylation across HELIX cohorts. **e**) Interaction plot illustrating differences across inv-8p23.1 genotypes in the association between cg08196601 (*TRMT9B*) methylation and parental smoking (N=1,009). **f**) Forest plot showing the meta-analysis effect estimates of the inv-8p23.1-parental smoking interaction on cg08196601 methylation across HELIX cohorts. Blue points and lines illustrate non-inverted homozygous (N/N), yellow illustrate heterozygous (N/I), and orange illustrate inverted homozygous (I/I) individuals. The error bar represents one standard deviation.

Another interesting result of our analysis relates to the methylation of the *TRMT9B* gene, also known as *C8orf79* or *KIAA1456*, a tRNA methyltransferase. The gene has been seen to associate with laryngotracheitis, an upper respiratory tract disease in chicken[43,44]. We observed that parental smoking during childhood significantly modulated the inversion-associated methylation of cg08196601 (β = -0.010, P = $1.3x10^{-10}$) (Figure 3e). The interaction of the inversion with maternal smoking during pregnancy was also associated with the methylation of cg08196601 (β = -0.020, P = $5.9x10^{-8}$). In addition, the methylation of cg26339990 was associated with the interaction of the inversion with outdoor PM2.5 (an air pollution exposure) during pregnancy (β = -0.003, P = $5.5x10^{-8}$). In the three cases, the non-inverted allele was associated with increased levels of methylation with the exposures. We observed that the heterogeneity across cohorts was not significant (P = 0.63) (Figure 3f). In line with these observations, the non-inverted allele for inv-8p23.1 has been found to associate with asthma[5] while parental smoking and exposure to high levels of PM2.5 during pregnancy or childhood increase the risk of respiratory diseases in children[45–47].

Discussion

Here, we show that the common human chromosomal inversions at 8p23.1, 16p11.2, and 17q21.31 have distinctive methylation patterns in blood across the inverted regions and that the early-life exposome modulates these patterns. We observed that during childhood approximately 10% of the CpG sites within the inverted regions +/- 1 Mb were significantly differentially methylated according to the inversion genotype. The amount of the differentially methylated CpG sites was high within the region and sharply decreased after the breakpoints, indicating the targeted effect of genomic inversions on DNA methylation. We could also identify the effects

of the inversions at prenatal stages in heart tissue, suggesting their relevant role during development even in utero. As such, inversions are early methylation quantitative loci for the genes they enclose. Our findings, therefore, add to other effects that inversions have on gene expression[8,13,14,48], derived from their genetic variability or from the displacement of regulatory elements near the breakpoints[10]. While individual CpG associations with the inversion may be due to the inversion or to local genetic variability in linkage with the inversion, our observations in the PC analysis reveal a spatial pattern given by the correlation of several CpG sites associations that fits the extension of the inversion. It is clear that the cause of such extended pattern along the affected sequence has been produced by the presence of the inversion, likely due to both the DNA reconfiguration and the accumulation of specific genetic variability along the segment that results from the suppression of recombination between inversion states.

We show that an important influence of inversions on phenotypes could be derived from the methylation patterns they support. Few previous studies have analyzed targeted methylation changes when studying a specific inversion or disease. We previously reported that the effect of inv-17q21.31 on colorectal disease-free survival is more likely mediated by DNA methylation than by gene expression[6]. Here, we document that the effect of inversions on methylation is strong along the inverted segment and already significant during early embryonic and fetal development in heart tissue DNA. One of the main stablished mechanisms underlying the influence of inversions on phenotypic traits and their pleiotropy is the suppression of recombination within the inverted sequence in heterozygotes. Allele combinations can thus be protected, leading to the generation and possible selection of specific haplotypes for each inversion state[10]. In addition, inversion breakpoints can disrupt coding regions or regulatory elements,

altering gene expression or generating novel transcripts with phenotypic consequences, including deleterious effects[15]. These effects likely play a role in the association of these three polymorphic inversions with complex diseases, like obesity[5,8], autoimmune diseases[49], or neurodegenerative disorders[50–52]. For these diseases with important environmental components, our results further suggest the additional role of inversion-associated methylation that is modifiable by environmental exposures.

Allele-specific methylation patterns in inversions can be caused or facilitated by their specific genetic variability and/or different chromatin structure. In our study, we removed probes with SNPs within 5bp distance and overall population frequency higher than 1%, ruling out technical and genetic variation as main contributors to the methylation differences. We observed that inversions at 8p23.1 and 17q21.31 were strongly characterized by their methylation patterns in the region. However, the effect was less strong for inv-16p11.2, which can be due to the higher number of haplotype groups supported by the inversion, that is two distinct haplotype groups in the standard allele and one in the inverted allele, and the fact that this inversion is smaller in size (0.45Mb versus 0.9Mb for inv-17q21.31 and almost 4Mb for inv-8p23.1)[8]. These specific effects on the methylation patterns could be mainly caused by differences in the three-dimensional (3D) DNA configuration for each allele[26], rendering some haplotypes more accessible to the different factors that could facilitate DNA methylation. This mechanism would explain how a recurrent but non-polymorphic inversion at Xq28 causing Hemophilia A has been associated with specific methylation changes[23] or how de novo inversions at 11p15.5 causing Beckwith-Wiedemann syndrome can be hypermethylated[24]. The possible correlation of inversion haplotypes with different 3D configurations and nuclear localization should be investigated in future studies.

We found that while the effects of the inversion on gene transcription and CpG methylation are widespread across the affected region with some overlap, the specific expression changes driven by inversion-association methylation need to be individually assessed. While the extended pattern of methylation across the inversion can be a consequence of the reconfiguration of the chromatin structure, gene expression may be more susceptible to the tissue and the local genetic variability in linkage with an inversion allele. In the case of 17q21 inversion, for instance, we found clear methylation patterns associated with inversion alleles, but no expression differences, which suggests that these methylation changes would have no relevant consequences in blood. By contrast, we also identified a relevant and specific mediator role by the methylation at promoters of *TUFM* and *SULT1A1* on the associations of their expressions with inv16p11.2. Remarkably, these are candidate genes in the association between inv-16p11.2 and the co-occurrence of asthma and obesity[8].

Previous studies have reported transcriptomic effects of inv-17q21.31 in blood only in genes with multiple copies[53,54]. This is a complex region with high variability in the gene copies within the inversion alleles, high homology between the genes with multiple copies, and low expression of the genes in blood[14,55]. This could explain the lack of *eQTL* effects of inv-17q21 in blood that we observed.

We have found that several methylation effects of inversions are modifiable by numerous environmental exposures, suggesting additional inversion-methylation effects to those driven by genetic variability. We observed that inversions significantly interacted with a wide range of exposures affecting DNA methylation across the inverted segments. Therefore, inversions are common copy-neutral polymorphisms that seem to be important contributors to gene-environment interactions, whose detection remains elusive in genomic and high dimensional exposure data[56–58]. We analyzed

data from an exposome study, covering a wide range of exposure families believed to affect children's development. The exposome data included environmental exposures but also exposures from the diet, urban exposome, and chemical compounds[31]. In total, we assessed 64 exposures (7 during pregnancy and 57 at 6-11 years of age) grouped in 12 families. We observed inversion interactions in most of the exposure families, most prominently in metals, diet, phenols, and organochlorines. Validation of these results and their consequences remain to be evaluated. Our results support the notion that inversions can change the way exposures affect a child's development by changing the genetic context. Carriers of genomic variants, such as these inversions that may affect the function of a set of genes in a specific direction, can be more susceptible to (or naturally protected against) disease or developmental disorders if exposed to a relevant environmental risk factor[59]. Thus, allele-specific methylation in response to different environmental factors could also contribute to the positive selection that has been documented for all three inversions in some human populations[8,12,60].

We found numerous significant inversion-exposure interactions on methylation levels in important genes which deserve further study. These include amongst others, Alzheimer's *MAPT* and its associations with copper[61], *MSRA*'s role in repairing oxidative damage to proteins and its relation with diet and parental smoking, and the oncogene *WNT3* and its relation to molybdenum and mercury. Here, we highlight three interactions with potential clinical interest and substantial support from previous studies. First, we observed the interaction of inv-8p23.1 with meat intake associated with *TDH* methylation levels. Remarkably, the inversion, the exposure, and the gene are independently associated with obesity in adults[5,39–41]. Our data revealed that non-inverted homozygous individuals, those with a higher risk of obesity, decreased methylation of two CpG sites

within *TDH* as meat intake increases. While further studies are needed to describe the role that this pseudogene plays in obesity during development, it is clear that these need to incorporate the effects of the inversion and its methylation status. In addition, clinical interventions of obesity aiming at managing meat intake should consider the methylation of the gene and the inversion genotype of individuals. Second, we observed that cg26020513 within *GATA4* was hypermethylated in blood when manganese exposure increased but only in non-inverted homozygous individuals. It is notable that the hypermethylation of cg26020513 has been strongly associated with congenital heart defects in fetuses[42], mutations in *GATA4* have been associated with cardiac septal defects[62], and manganese toxicity in heart tissue is well documented[63]. The inversion also interacted with other relevant exposures on *GATA4* methylation including mercury, with reported effects in heart rate variability in children[64], diethylphosphate, Mediterranean diet, and PCB 138. Therefore, the extent to which the inversion status can protect against the positive association between these exposures and *GATA4* methylation deserves further scrutiny. Third, we observed that the effects of tobacco smoke (during pregnancy or in childhood) and air pollution (outdoor PM2.5 exposure) on *TRMT9B* methylation changed depending on the inv-8p23.1 genotype. Since these two exposures increase the risk of respiratory diseases[45–47] and *TRMT9B* is a gene associated with an upper respiratory tract disease[43,44], our results suggest a likely role of the gene in the association between inv-8p23.1 and asthma[5].

To the best of our knowledge, this is the first study to systematically assess the methylation landscape within three common human inversions and its interaction with the exposome. We have shown that genomic inversions are associated with the methylation of the CpG sites within the inversion region

and that this association is modulated by a wide range of environmental exposures during childhood.

Methods

Study population

The Human Early Life Exposome (HELIX) project[36] comprises a total of 1,301 mother-child pairs from six birth cohorts in Europe: BIB (Born in Bradford; the United Kingdom)[65], EDEN (Etude des Déterminants pré et postnatals du développement et de la santé de l'Enfant; France)[66], INMA-SAB (Infancia y Medio Ambiente; Spain; subcohort Sabadell)[67], KANC (Kaunas cohort; Lithuania)[68], MoBa (The Norwegian Mother, Father and Child Cohort study; Norway)[69], and Rhea (Greece)[70]. These mother-child pairs participated in a common, completely harmonized, follow-up examination between December 2013 and February 2016, when children were between 6-11 years old[71]. The main goal of this project was to implement exposure assessment and biomarker methods to characterize early-life exposure to multiple environmental factors and associate these with omics biomarkers and child health outcomes. For these same children, multi-omics molecular phenotyping was performed, including measurement of blood DNA methylation (450K, Illumina), blood gene expression (HTA v2.0, Affymetrix), blood miRNA expression (SurePrint Human miRNA rel 21, Agilent), plasma proteins (Luminex), serum metabolites (AbsoluteIDQ p180 kit, Biocrates), urinary metabolites (1H NMR spectroscopy), and DNA microarray (Chemagen kit, Perkin Elmer). All studies received approval from the ethics committees of the centers involved and written informed consent was obtained from all participants.

Molecular phenotypes

Inversion genotype data

DNA was obtained from buffy coat collected in EDTA tubes at 6-11 years of age. Briefly, DNA was extracted using the Chemagen kit (Perkin Elmer) in batches of 12 samples. Samples were extracted by cohort and following their position in the original boxes. DNA concentration was determined in a NanoDrop 1000 UV-Vis Spectrophotometer (ThermoScientific) and with Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies). Genome-wide genotyping was performed using the Infinium Global Screening Array (GSA) MD version 1 (Illumina) at the Human Genomics Facility (HuGe-F), Erasmus MC (www.glimdna.org). Genotype calling was done using the GenTrain2.0 algorithm based on a custom clusterfile for 692,367 variants implemented in the GenomeStudio software. Annotation was done with the GSAMD-24v1-0_20011747_A4 manifest, SNP coordinates were reported on human reference GRCh37 and Source strand (Forward strand report in GenomeStudio). The initial dataset consisted of 1,397 samples and 692,367 variants. Samples with discordant sex, duplicated, contaminated (high heterozygosity), and relatives (IBD >0.185) were filtered out. SNPs with variant call rate <95%, minimum allele frequency <1%, and HWE *P*-value (1x10-6) were excluded. Major population ancestry groups were estimated using *Peddy*[37] and only individuals of European ancestry were kept in the analysis. The final dataset consisted of 1,009 samples and 509,344 SNP variants. From this dataset, we selected inversions that could be genotyped with *scoreInvHap* and had more than 10 CpG sites in the inversion region: inv-8p23.1, inv-16p11.2, and inv-17q21.31 (Table 2 and Supplementary Table 2-3).

DNA methylation

The DNA was obtained using the same methodology as for genetics data. DNA methylation was assessed using the Infinium Human Methylation 450 beadchip (Illumina), following the manufacturer's protocol. *Minfi* R package[72] was used for the pre-processing of DNA methylation data. *MethylAid* package[73] was employed to perform the first quality control of the data. Probes with low call rates were filtered following the guidelines of Lehne et al.[74] The functional normalization method was further applied, including Noob background subtraction and dye-bias correction[75]. Several quality control checks were performed: sex consistency using the *shinyMethyl* package[76], consistency of duplicates, and genetic consistency for the samples that had genome-wide genotypic data. Duplicated samples and control samples were removed as well as probes that measure methylation levels at non-CpG sites[77]. Probes that cross-hybridize were excluded. Moreover, we used InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation to filter probes where 30bp 3'-subsequence of the probe is non-unique, probes with INDELs, probes with extension base inconsistent with specified color channel (type-I) or CpG (type-II) based on mapping, probes with a SNP in the extension base that causes a color channel switch from the official annotation, and probes where 5bp 3'-subsequence overlap with any of the SNPs with global population frequency higher than 1%. Consequently, the number of CpG probes analyzed was 371,533, initially available for 1,192 subjects. We then used *Combat* algorithm to remove the batch effects supported by the slide. Methylation levels were expressed as beta values (average methylation levels for an individual, between 0 for a never methylated CpG site and 1 for an always-methylated CpG site) and CpG sites were annotated to genes by Illumina HM450 manifest file (version 1.2). We discarded the subjects without inversion status data and without European ancestry based on

genomic data, resulting in 1,009 individuals for the analysis. For each inversion, we selected the CpG sites contained in the inversion region +/-1 Mb, resulting in 848 CpG sites for inv-8p23.1, 401 for inv-16p11.2, and 666 for inv-17q21.31 (Table 2 and Supplementary Table 2). Blood cell type proportions were estimated from methylation data according to Houseman et al. algorithm[78] and Reinius reference panel[79].

Table 2 | Characteristics of HELIX data relating 3 common polymorphic inversions in humans. The table shows the length in kb, the mapping coordinates hg19 +/-1Mb, the frequency of all the inversions obtained from *scoreInvHap*[11], and the number of samples and features used in transcriptome and methylome analysis for each inversion. *The allele in the reference genome is the least frequent in the population.

Genomic inversion	Length (kb)	Inversion region +/- 1Mb	Inversion frequency (%)	Omics	Number of samples	Number of features
8p23.1	3,924.86	chr8:7055789-12980649	57.95	Methylome	1009	848
				Transcriptome	926	83
16p11.2	364.17	chr16:27424774-29788943	34.49	Methylome	1009	401
				Transcriptome	926	58
17q21.31	710.89	chr17:42661775-45372665	23.96	Methylome	1009	666
				Transcriptome	926	61

Gene expression

At the period of clinical examination that took place when children were between 6 and 11 years old, RNA was extracted from whole blood collected in Tempus tubes. Samples with RIN >5 were considered. Gene expression was assessed using the GeneChip® Human Transcriptome Array 2.0 (HTA 2.0) (Affymetrix, USA) at the University of Santiago de Compostela (USC, Spain), following the manufacturer's protocol. Samples were randomized and balanced by sex and cohort within each batch. Data were normalized at the gene level with the GCCN (SST-RMA) algorithm, and batch effects and blood cell type composition were controlled with two surrogate variable analysis (SVA) methods, *isva*[80] and *SmartSVA*[81], during the differential

expression analyses. Gene expression values were log2 transformed, and annotation of transcript clusters (TCs) to genes was done with NetAffx annotation (version 36). Genes without Gene Symbol annotation or with call rate <20% were removed, restricting to 25,255 genes. From this number of genes, we selected those within the inversion regions +/- 1Mb (inv-8p23.1: 83 genes; inv-16p11.2: 58 genes; inv-17q21.31: 61 genes). From a total of 1,158 subjects that had transcriptomic data, we selected individuals with European ancestry (based on genomic data) who had available inversion status data and cell type proportions assessed from methylation data, resulting in a total of 790 subjects (Table 2 and Supplementary Table 1).

Exposome assessment

The assessment of the exposome has been previously published[82]. In our study, we included 7 exposures assessed during pregnancy and 57 exposures assessed during childhood at age 6-11y (Supplementary Data 4). These 64 exposures were selected from the entire exposome dataset according to the number of missing values they had. We did not include exposures that had more than 10% of missing in the whole dataset or with more than 20% missing in one or more cohorts. We also excluded exposures whose levels were not present in all cohorts. Third, we selected the most representative exposures within each family.

The pregnancy exposome consists of 7 exposures, including outdoor PM2.5, normalized difference vegetation index (NDVI), 4 PFASs, and maternal smoking during pregnancy. The postnatal exposome was divided into 12 exposure families: outdoor air pollution (2), building environment (1), diet (6), metals (9), natural spaces (1), organochlorines – OCs (8), organophosphate pesticides – OP pesticides (5), polybrominated diphenyl ethers – PBDEs (2), perfluorinated alkylated substances – PFAS (5),

phenols (7), phthalates (10), and second-hand exposure to tobacco smoke (1) (Figure 2a). Metals, OCs, OP Pesticides, PBDEs, PFASs, phenols, and phthalates were assessed by biomarkers in children at the time of the clinical examination, from a pool of two urine samples or one serum sample[83]. Air pollution, natural spaces, and building environment quantification were assessed during the year before child examination or during pregnancy by environmental geographic information systems (GIS). Tobacco smoke and diet were evaluated by questionnaires. Missing values for all exposures were imputed using the method of chained equations[84], as described in detail elsewhere[82]. Most exposure variables were transformed as described in Supplementary Data 4.

Fetal heart tissue samples

Human fetal samples from 40 fetuses of terminated pregnancies due to a major congenital heart defect (gestational age 21-22 weeks in all cases) were obtained from Biobanc Hospital Universitari Vall d'Hebron (HUVH) in a related project addressed to define the genetic and epigenetic basis of congenital heart defects[38]. Informed consent was obtained from parents and the study was approved by the institutional ethics committee. Heart tissue DNA was obtained following necropsy using standard procedures, whole genome sequencing was performed at Centogene, and DNA methylation was measured with Infinium MethylationEPIC[38].

After quality control, one sample was discarded (Supplementary Table 4). During the pre-processing of methylation data, probes with a single-nucleotide polymorphism (SNP) with overall population frequency higher than 1% based on InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation were removed. Selecting the CpG sites within the inversion region +/- 1Mb, we analyzed 898 CpG sites from inv-8p23.1, 409 from inv-16p11.2, and 698 from inv-17q21.31.

Statistics and Reproducibility

Genome-wide analysis

Differential methylation analyses were performed using *MEAL* Bioconductor's package[85]. We performed a differential mean analysis (DMA) on inversion genotypes using the function *runDiffMeanAnalysis* that calls *limma*[86]. Based on a priori knowledge, we adjusted all the regression models by sex, age, population stratification (using the first 10 principal components of the GWAS that highly correlated with cohort), and cell type (Supplementary Table 1-2). To correct for the variance between cohorts, we performed this analysis for each cohort separately and we meta-analyzed the results using the function *metagen* from *meta* package[87]. For each inversion, in each cohort, we fitted models

$$E_j = \alpha_j + \beta_{jk} I_k + \Sigma_r \gamma_r C_r + \varepsilon_j \qquad (1)$$

where E_j is the methylation or expression level vector across individuals at probe j, I_k are the individuals' genotypes for inversion k (8p23.1, 16p11.2, 17q21.31), C_r is the r covariate and its effect γ_r, and ε_j is the noise that follows the distribution of methylation or expression levels with mean 0. β_{jk} is the effect of interest measuring the effect of the inversion. The β_{jk} were then meta-analyzed across cohorts. *P*-values derived from the meta-analyses were corrected for multiple comparisons for the number of probes using Bonferroni's correction. The inflation or deflation of *P*-values across the methylome or transcriptome was tested with Q-Q plots.

Exposome-wide interaction analysis

Based on the genome-wide analysis, the same functions were implemented for the exposome-wide interaction analysis. In this case, the effect of interest was the inversion-exposure interaction in the model

$$E_j = \alpha_j + \beta_{jik}\,(X_i \times I_k) + \Sigma_r \gamma_r C_r + \varepsilon_j \qquad (2)$$

where X_i is the level of exposure i across individuals. β_{jik} is the effect of interest given by the exposure-inversion interaction. In this case, the covariates also included exposure i, the inversion genotypes, maternal education level, and child body mass index (BMI). *P*-values were corrected for multiple comparisons across CpG sites and exposures using Bonferroni's correction. The inflation or deflation of *P*-values across the methylome was tested with Q-Q plots.

References

1. Martínez-Fundichely, A. et al. InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res. 42, D1027–D1032 (2013).

2. Kirkpatrick, M. & Barton, N. Chromosome inversions, local adaptation and speciation. Genetics 173, 419–434 (2006).

3. Sturtevant, A. H. & Beadle, G. W. The Relations of Inversions in the X Chromosome of Drosophila Melanogaster to Crossing over and Disjunction. Genetics 21, 554–604 (1936).

4. Cáceres, A. & González, J. R. Following the footprints of polymorphic inversions on SNP data: From detection to association tests. Nucleic Acids Res. 43, [e53] (2015).

5. González, J. R. et al. Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases. Am. J. Hum. Genet. 106, 846–858 (2020).

6. Ruiz-Arenas, C., Cáceres, A., Moreno, V. & González, J. R. Common polymorphic inversions at 17q21.31 and 8p23.1 associate with cancer prognosis. Hum. Genomics 13, [57] (2019).

7. Tantisira, K. G., Lazarus, R., Litonjua, A. A., Klanderman, B. & Weiss, S. T. Chromosome 17: Association of a large inversion polymorphism with corticosteroid response in asthma. Pharmacogenet. Genomics 18, 733–737 (2008).

8. González, J. R. et al. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. Am. J. Hum. Genet. 94, 361–372 (2014).

9. Luciano, M. et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. Nat. Genet. 50, 6–11 (2018).

10. Puig, M., Casillas, S., Villatoro, S. & Cáceres, M. Human inversions and their functional consequences. Brief. Funct. Genomics 14, 369–379 (2015).

11. Ruiz-Arenas, C. et al. scoreInvHap: Inversion genotyping for genome-wide association studies. PLoS Genet. 15, [e1008203] (2019).

12. Giner-Delgado, C. et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nat. Commun. 10, [4222] (2019).

13. Salm, M. P. A. et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res. 22, 1144–1153 (2012).

14. de Jong, S. et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics 13, 458 (2012).

15. Lakich, D., Kazazian, H. H., Antonarakis, S. E. & Gitschier, J. Inversions disrupting the factor VIII gene are a common cause of severe haemophilia A. Nat. Genet. 5, 236–241 (1993).

16. Jaarola, M., Martin, R. H. & Ashley, T. Direct evidence for suppression of recombination within two pericentric inversions in humans: A new sperm-FISH technique. Am. J. Hum. Genet. 63, 218–224 (1998).

17. Ruiz-Arenas, C. et al. Identifying chromosomal subpopulations based on their recombination histories advances the study of the genetic basis of phenotypic traits. Genome Res. 31, 1802–1814 (2020).

18. Phillips, T. The Role of Methylation in Gene Expression | Learn Science at Scitable. Nat. Educ. 1, 116 (2008).

19. Suzuki, M. M. & Bird, A. DNA methylation landscapes: Provocative insights from epigenomics. Nature Reviews Genetics vol. 9 465–476 (2008).

20. Métivier, R. et al. Cyclical DNA methylation of a transcriptionally active promoter. Nature 452, 45–50 (2008).

21. Jaffe, A. E. et al. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. Int. J. Epidemiol. 41, 200–209 (2012).

22. Shi, X. et al. Association of CNVs with methylation variation. npj Genomic Med. 5, [41] (2020).

23. Jamil, M. A. et al. F8 inversions at Xq28 causing hemophilia a are associated with specific methylation changes: Implication for molecular epigenetic diagnosis. Front. Genet. 10, 508 (2019).

24. Smith, A. C. et al. Maternal gametic transmission of translocations or inversions of human chromosome 11p15.5 results in regional DNA hypermethylation and downregulation of CDKN1C expression. Genomics 99, 25–35 (2012).

25. Robertson, K. D. DNA methylation and chromatin – unraveling the tangled web. Oncogene 2002 2135 21, 5361–5379 (2002).

26. Shanta, O. et al. The effects of common structural variants on 3D chromatin structure. BMC Genomics 21, 1–10 (2020).

27. Marsit, C. J. Influence of environmental exposure on human epigenetic regulation. Journal of Experimental Biology vol. 218 71–79 (2015).

28. Bollati, V. & Baccarelli, A. Environmental epigenetics. Heredity vol. 105 105–112 (2010).

29. Stein, R. A. Epigenetics and environmental exposures. J. Epidemiol. Community Health 66, 8–13 (2012).

30. Wild, C. P. Complementing the Genome with an '"Exposome"': The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. 14, 1847–50 (2005).

31. Miller, G. W. & Jones, D. P. The nature of nurture: Refining the definition of the exposome. Toxicol. Sci. 137, 1–2 (2014).

32. Hing, B., Gardner, C. & Potash, J. B. Effects of negative stressors on DNA methylation in the brain: Implications for mood and anxiety disorders. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics vol. 165 541–554 (2014).

33. Hunter, R. G. & McEwen, B. S. Stress and anxiety across the lifespan: Structural plasticity and epigenetic regulation. Epigenomics vol. 5 177–194 (2013).

34. Teh, A. L. et al. The effect of genotype and in utero environment on interindividual variation in neonate DNA methylomes. Genome Res. 24, 1064–1074 (2014).

35. Law, P. P. & Holland, M. L. DNA methylation at the crossroads of gene and environment interactions. Essays in Biochemistry vol. 63 717–726 (2019).

36. Vrijheid, M. et al. The human early-life exposome (HELIX): Project rationale and design. Environmental Health Perspectives vol. 122 535–544 (2014).

37. Pedersen, B. S. & Quinlan, A. R. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am. J. Hum. Genet. 100, 406–413 (2017).

38. Ruiz-Arenas, C. A multi-omics approach improves diagnosis in major isolated congenital heart disease. ASHG Virtual Meet. (2020).

39. Schlauch, K. A. et al. A Comprehensive Genome-Wide and Phenome-Wide Examination of BMI and Obesity in a Northern Nevadan Cohort. G3 10, 645–664 (2020).

40. MH, R., A, S.-A., PJ, S. & L, A. Is there a relationship between red or processed meat intake and obesity? A systematic review and meta-analysis of observational studies. Obes. Rev. 15, 740–748 (2014).

41. You, W. & Henneberg, M. Meat consumption providing a surplus energy in modern diet contributes to obesity prevalence: an ecological analysis. BMC Nutr. 2016 21 2, 1–11 (2016).

42. Serra-Juhé, C. et al. DNA methylation abnormalities in congenital heart disease. Epigenetics 10, 167–177 (2015).

43. JA, C. et al. Methylome Analysis in Chickens Immunized with Infectious Laryngotracheitis Vaccine. PLoS One 10, [e0100476] (2015).

44. J, L., WG, B. & BW, K. Genome-wide host responses against infectious laryngotracheitis virus vaccine infection in chicken embryo lung cells. BMC Genomics 13, [143] (2012).

45. M, S., A, H.-T., C, G., RM, H. & TH, S. Smoking and asthma. J. Am. Board Fam. Med. 24, 313–322 (2011).

46. A, Z. Maternal smoking in pregnancy and its influence on childhood asthma. ERJ open Res. 2, 00042–02016 (2016).

47. L, Q. et al. The impact of PM2.5 on lung function in adults with asthma. Int. J. Tuberc. Lung Dis. 24, 570–576 (2020).

48. Puig, M. et al. Functional Impact and Evolution of a Novel Human Polymorphic Inversion That Disrupts a Gene and Creates a Fusion Transcript. PLoS Genet. 11, [e1005495] (2015).

49. Namjou, B. et al. The Effect of Inversion at 8p23 on BLK Association with Lupus in Caucasian Population. PLoS One 9, [e115614] (2014).

50. Webb, A. et al. Role of the tau gene region chromosome inversion in progressive supranuclear palsy, corticobasal degeneration, and related disorders. Arch. Neurol. 65, 1473–1478 (2008).

51. Myers, A. J. et al. The H1c haplotype at the MAPT locus is associated with Alzheimer's disease. Hum. Mol. Genet. 14, 2399–2404 (2005).

52. Setó-Salvia, N. et al. Dementia risk in parkinson disease: Disentangling the role of MAPT Haplotypes. Arch. Neurol. 68, 359–364 (2011).

53. Degenhardt, F. et al. New susceptibility loci for severe COVID-19 by detailed GWAS analysis in European populations. medRxiv 9, 2021.07.21.21260624 (2021).

54. Puig, M. et al. Determining the impact of uncharacterized inversions in the human genome by droplet digital PCR. Genome Res. 30, 724–735 (2020).

55. Steinberg, K. M. et al. Structural diversity and African origin of the 17q21.31 inversion polymorphism. Nat. Genet. 44, 872–880 (2012).

56. Li, J., Li, X., Zhang, S. & Snyder, M. Gene-Environment Interaction in the Era of Precision Medicine. Cell vol. 177 38–44 (2019).

57. Assary, E., Vincent, J. P., Keers, R. & Pluess, M. Gene-environment interaction and psychiatric disorders: Review and future directions. Seminars in Cell and Developmental Biology vol. 77 133–143 (2018).

58. Wu, M., Zhang, Q. & Ma, S. Structured gene-environment interaction analysis. Biometrics 76, 23–35 (2020).

59. Manuck, S. B. & McCaffery, J. M. Gene-environment interaction. Annual Review of Psychology vol. 65 41–70 (2014).

60. Stefansson, H. et al. A common inversion under selection in Europeans. Nat. Genet. 37, 129–137 (2005).

61. Bagheri, S., Squitti, R., Haertlé, T., Siotto, M. & Saboury, A. A. Role of Copper in the Onset of Alzheimer's Disease Compared to Other Metals. Front. Aging Neurosci. 9, 446 (2018).

62. Yang, Y.-Q. et al. Mutation spectrum of GATA4 associated with congenital atrial septal defects. Arch. Med. Sci. 9, 976 (2013).

63. Jiang, Y. & Zheng, W. Cardiovascular Toxicities Upon Manganese Exposure. Cardiovasc. Toxicol. 5, 345 (2005).

64. Genchi, G., Sinicropi, M. S., Carocci, A., Lauria, G. & Catalano, A. Mercury exposure and heart diseases. International Journal of Environmental Research and Public Health vol. 14 74 (2017).

65. Wright, J. et al. Cohort profile: The born in bradford multi-ethnic family cohort study. Int. J. Epidemiol. 42, 978–991 (2013).

66. 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. 45, 353–363 (2016).

67. Guxens, M. et al. Cohort profile: The INMA-INfancia y Medio Ambiente-(environment and childhood) project. Int. J. Epidemiol. 41, 930–940 (2012).

68. Grazuleviciene, R. et al. Surrounding greenness, proximity to city parks and pregnancy outcomes in Kaunas cohort study. Int. J. Hyg. Environ. Health 218, 358–365 (2015).

69. Magnus, P. et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). Int. J. Epidemiol. 45, 382–388 (2016).

70. Chatzi, L. et al. Cohort Profile: The Mother-Child Cohort in Crete, Greece (Rhea Study). Int. J. Epidemiol. 46, 1392–1393 (2017).

71. Vrijheid, M. et al. Environmental Exposures and Childhood Obesity: An Exposome Analysis. in ISEE Conference Abstracts (Environmental Health Perspectives, 2018). doi:10.1289/isesisee.2018.o02.01.24.

72. 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).

73. Van Iterson, M. et al. MethylAid: Visual and interactive quality control of large Illumina 450k datasets. Bioinformatics 30, 3435–3437 (2014).

74. Lehne, B. et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 16, 37 (2015).

75. Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. Nucleic Acids Res. 41, [e90] (2013).

76. Fortin, J. P., Fertig, E. & Hansen, K. shinyMethyl: Interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Research 3, 175 (2014).

77. Jang, H. S., Shin, W. J., Lee, J. E. & Do, J. T. CpG and non-CpG methylation in epigenetic gene regulation and brain function. Genes vol. 8 2–20 (2017).

78. Houseman, E. A. et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinforma. 2012 131 13, 1–16 (2012).

79. 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).

80. Teschendorff, A. E., Zhuang, J. & Widschwendter, M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics 27, 1496–1505 (2011).

81. Chen, J. et al. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. BMC Genomics 18, 413 (2017).

82. Tamayo-Uria, I. et al. The early-life exposome: Description and patterns in six European countries. Environ. Int. 123, 189–200 (2019).

83. LS, H. et al. In-utero and childhood chemical exposome in six European mother-child cohorts. Environ. Int. 121, 751–763 (2018).

84. IR, W., P, R. & AM, W. Multiple imputation using chained equations: Issues and guidance for practice. Stat. Med. 30, 377–399 (2011).

85. Ruiz-Arenas, C. & Gonzalez, J. R. MEAL: Perform methylation analysis. R package version 1.22.0. (2019).

86. Ritchie, M. E. et al. limma powers differential expression analyses for {RNA}-sequencing and microarray studies. Nucleic Acids Res. 43, [e47] (2015).

87. Balduzzi, S., Rücker, G. & Schwarzer, G. How to perform a meta-analysis with R: a practical tutorial. Evid. Based. Ment. Health 22, 153–160 (2019).

Code availability

Any custom code or software used in our analysis is available at https://doi.org/10.5281/zenodo.6417926
(URL: https://zenodo.org/badge/latestdoi/296552532).

Data availability

The HELIX data warehouse has been established as an accessible resource for collaborative research involving researchers external to the project. Access to HELIX data is based on approval by the HELIX Project Executive Committee and by the individual cohorts. Further details on the content of the data warehouse (data catalog) and procedures for external access are described on the project website (http://www.projecthelix.eu/index.php/es/data-inventory). The data used in this analysis are not available for replication because specific approvals from HELIX Project Executive Committee and the University of Southern California Institutional Review Board must be obtained to access them. Source data underlying Fig 2a, 3a, and 3e is available in Supplementary Data 7.

Acknowledgements

We are grateful to all the participating children, parents, practitioners, and researchers in the six countries who took part in this study. The study has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 308333 (HELIX project); and the H2020-EU.3.1.2. - Preventing Disease Programme under grant agreement no 874583 (ATHLETE project). The HELIX genotyping was supported by the projects PI17/01225 and PI17/01935, funded by the Instituto de Salud Carlos III and co-funded by European Union (ERDF, "A

way to make Europe") and the Centro Nacional de Genotipado-CEGEN (PRB2-ISCIII).

BiB received core infrastructure funding from the Wellcome Trust (WT101597MA) and a joint grant from the UK Medical Research Council (MRC) and Economic and Social Science Research Council (ESRC) (MR/N024397/1). INMA-SAB data collections were supported by grants from the Instituto de Salud Carlos III, CIBERESP, and the Generalitat de Catalunya-CIRIT. KANC was funded by the grant of the Lithuanian Agency for Science Innovation and Technology (6-04-2014_31V-66). The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. The Rhea project was financially supported by European projects (EU FP6-2003-Food-3-NewGeneris, EU FP6. STREP Hiwate, EU FP7 ENV.2007.1.2.2.2. Project No 211250 Escape, EU FP7-2008-ENV-1.2.1.4 Envirogenomarkers, EU FP7-HEALTH-2009- single stage CHICOS, EU FP7 ENV.2008.1.2.1.6. Proposal No 226285 ENRIECO, EU- FP7- HEALTH-2012 Proposal No 308333 HELIX), and the Greek Ministry of Health (Program of Prevention of obesity and neurodevelopmental disorders in preschool children, in Heraklion district, Crete, Greece: 2011-2014; "Rhea Plus": Primary Prevention Program of Environmental Risk Factors for Reproductive Health, and Child Health: 2012-15).

This research has received funding from the Spanish Ministry of Education, Innovation and Universities, the National Agency for Research and the Fund for Regional Development (RTI2018-100789-B-I00), MaratóTV3 (2015-3230), the Spanish Ministry of Science and Innovation through the "Centro de Excelencia Severo Ochoa 2019-2023 (CEX2018-000806-S) and Maria de Maeztu (MDM-2014-0370)" Programs, and support from the Generalitat de Catalunya through the CERCA and Consolidated Research

Group (2017SGR01974) Programs. NC and JU are supported by Spanish regional program PERIS (Ref.: SLT017/20/000061 and SLT017/20/000119, respectively), granted by Departament de Salut de la Generalitat de Catalunya. We thank Pau Bosch Castro for designing and creating the featured image.

Author contributions

JRG conceived the study and supervised genomic inversion analyses. JRG, AC, and LAP-J designed the analysis. LB-D performed genomic inversion calling and NC-G the statistical analyses. MV coordinates the HELIX project, JU is the data manager, and LM is the scientific coordinator. MB, JW, RS, MC, JRG, and MV designed the omics study in HELIX. The following authors participated in omics data acquisition and quality control: GE (genomics), MB (transcriptomics and DNA methylation), AnC (DNA methylation), MJN (exposome), and CT (exposome). JW, RG, MV, RS, LC, and CT are the PIs of the cohorts. TY, SA, MC, JL, NS, and KG participated in sample and data acquisition. CR-A performed inversion-methylation analyses in heart tissue. LAP-J coordinated the study of heart tissue in CHD. NC-G and AC co-wrote the original draft of the paper and JRG, LAP-J, MB, CR-A, and LB-D contributed to review and edit the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper online.

Correspondence and requests for materials should be addressed to JR.G.

5.3. Manuscript 3

A prenatal environment of female protection against childhood obesity is associated with sex differences in neurodevelopment and adult academic achievement

Cáceres A [1], **Carreras-Gallo N [1]**, Andrusaityte S, Bustamante M, Carracedo A, Chatzi L, Dwaraka VB, Grazuleviciene R, Gutzkow KB, Lepeule J, Maitre L, Mendez T, Nieuwenhuijsen M, Slama R, Smith R, Stratakis N, Thomsen C, Urquiza J, Went H, Wright J, Yang T, Casas M, Vrijheid M, González JR

[1] These authors contributed equally.

A prenatal environment of female protection against childhood obesity is associated with sex differences in neurodevelopment and adult academic achievement. Submitted to *BMC Medicine*. IF: 11.8. Position: Q1-D2

ARTICLE

A prenatal environment of female protection against childhood obesity is associated with sex differences in neurodevelopment and adult academic achievement

Alejandro Cáceres[1,2,3,†,*], Natàlia Carreras-Gallo[1,†], Sandra Andrusaityte[4], Mariona Bustamante[1,2,5,6], Ángel Carracedo[7,8], Leda Chatzi[9], Varun B. Dwaraka[10], Regina Grazuleviciene[4], Kristine Bjerve Gutzkow[11], Johanna Lepeule[12], Léa Maitre[1,2,5], Tavis L. Mendez[10], Mark Nieuwenhuijsen[1,2,5], Remy Slama[12], Ryan Smith[10], Nikos Stratakis[1], Cathrine Thomsen[11], José Urquiza[1,2,5], Hannah Went[10], John Wright[13], Tiffany Yang[13], Maribel Casas[1,2,5], Martine Vrijheid[1,2,5], Juan R. González[1,2,14,*]

Abstract

Background: Obesity and neurodevelopmental delay are complex traits that often co-occur and differ between boys and girls. Prenatal exposures are believed to influence children's obesity, but it is unknown whether exposures of pregnant mothers can confer a different risk of obesity between sexes, and whether they can affect neurodevelopment and adult academic achievement.

Methods: We analyzed data from 1,044 children from the HELIX project, comprising 93 exposures during pregnancy, and clinical, neuropsychological, and methylomic data during childhood (5-11 years). Using exposome-wide interaction analyses, we identified prenatal exposures with the highest sexual dimorphism in obesity risk, which were used to create a multiexposure profile. We applied causal random forest to classify individuals into two environments: E1 and E0. E1 consists of a specific combination of exposure levels where girls have significantly less risk of obesity than boys as compared to E0, which consists of the remaining combination of exposure levels. We investigated whether E1 had a lower female-

[1] Instituto de Salud Global de Barcelona (ISGlobal), Barcelona 08003, Spain, [2] Centro de Investigación Biomédica en Red en Epidemiología y Salud Pública (CIBERESP), Spain, [3] Department of Mathematics, Escola d'Enginyeria de Barcelona Est (EEBE), Universitat Politècnica de Catalunya, Barcelona 08019, Spain, [4] Department of Environmental Science, Vytautas Magnus University, 44248 Kaunas, Lithuania, [5] Department of Health and Experimental Sciences, Universitat Pompeu Fabra (UPF), Barcelona, Spain, [6] Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain, [7] Medicine Genomics Group, Centro de Investigación Biomédica en Red Enfermedades Raras (CIBERER), University of Santiago de Compostela, CEGEN-PRB3, Santiago de Compostela, Spain, [8] Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Servicio Gallego de Salud (SERGAS), Santiago de Compostela, Galicia, [9] Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, USA, [10] TruDiagnostic, Lexington, KY, United States, [11] Department of Environmental Health, Norwegian Institute of Public Health, 0456, Oslo, Norway, [12] Institut national de la santé et de la recherche médicale (Inserm) and Université Grenoble-Alpes, Institute for Advanced Biosciences (IAB), Team of Environmental Epidemiology applied to Reproduction and Respiratory Health, Grenoble, France, [13] Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK, [14] Department of Mathematics, Universitat Autònoma de Barcelona, Bellaterra (Barcelona) 08193, Spain. † These authors contributed equally. * Emails: alejandro.caceres@isglobal.org and juranr.gonzalez@isglobal.org

associated risk of neurodevelopmental delay than E0. We performed an epigenome-wide association study between the environments and assessed whether the methylation profile tagged to E1 was associated with sex differences in adult obesity and academic achievement in a large cohort (N=3,259).

Results: We observed that E1 was defined by the combination of low dairy consumption, low cotinine levels in blood, low facility richness, and the presence of green spaces during pregnancy ($OR_{interaction}$=0.070, *P*=2.59×10-5). E1 was also associated with a lower risk of neurodevelopmental delay in neuropsychological tests of non-verbal intelligence ($OR_{interaction}$=0.42, *P*=0.047), working memory ($OR_{interaction}$=0.31, *P*=0.02) and with methylation probes enriched with several neurodevelopmental functions. The methylation profile linked to E1 was significantly associated with sex differences in adult academic achievement ($OR_{interaction}$=1.58, *P*=0.008) and not obesity.

Conclusions: The risk of obesity can be different for boys and girls in certain prenatal environments. We identified an environment defined as a combination of four exposure levels that protects girls from obesity. The environment was associated with a lower risk of neurodevelopmental delay and higher academic achievement in adult life. The combination of single exposures into multiexposure profiles using causal inference can help determine populations at risk.

Keywords: prenatal environment, sexual dimorphism, childhood obesity, neurodevelopment, DNA methylation, causal inference, multiexposure profile

Introduction

Boys and girls develop differently. For instance, their immune response to infections differs from an early age, their brains grow at different rates, and the prevalence of numerous common diseases, like obesity, is also different [1–3]. Boys are more susceptible to obesity than girls [3]. Given the contrasting paths of development, it is remarkable that biomedical studies typically consider sex as a confounder rather than the main effect or an effect modifier [4]. Exposome studies, in particular, are characterized by the acquisition of massive amounts of data at individual and population levels [5]. A crucial goal of these studies is to inform the likely conditions for which a given public health intervention would be optimal, such that the best intervention is applied at the right time to the right population [6]. However, as the main difference between individuals is sex, exposome studies aiming at improving precision medicine and precision public health cannot do without considering how environmental risk factors affect sexual dimorphism in development and disease.

From a mechanistic context, studying the factors that increase sexual dimorphic outcomes of disease can offer important insights into its etiology and comorbidities, and inform of possible interventions and targeted treatments. Important advancements have been made in studying sex-related risk factors for diseases like cancer, Alzheimer's, and autoimmune diseases [7]. However, a relevant component of these age-related diseases is hormonal regulation. Studying sex differences in preteens offers not only the opportunity for identifying targeted treatments for early-age illnesses but also to explore disease mechanisms unlikely influenced by sex hormones that may also onset early in life. Previous research has, for instance, underlined that maternal factors during pregnancy can affect

disease outcomes later in life [8] and, therefore, motivates the question of which pregnancy factors may promote later sexual dimorphism in disease.

Environmental exposures likely orchestrate environments that are more toxic to one sex than to the other one. However, methods to determine such multiple-exposure environments are not readily available. We have developed a method of causal modeling, based on causal random forest, that can determine profiles of multiple exposures that are associated with high sexual dimorphism [9]. Here, we aimed to adapt our method to determine which combination of prenatal exposures can produce an environment where girls are more protected from obesity than boys during the preteen years. Furthermore, obesity in children is associated with lower cognitive function, particularly inhibitory control and working memory, critical for academic achievement [10]. Obesity often co-occurs with neurodevelopmental disorders, particularly in boys [11]. Therefore, we also evaluated whether the environment of high sexual dimorphism in obesity also shows a significant sexual dimorphism in non-verbal intelligence, working memory, attention, and ADHD.

Finally, we investigated whether a methylation profile may be associated with the protective environment since many exposures during pregnancy are associated with specific methylation profiles [12]. We then used methylation data from adults for targeting individuals who likely belonged to this profile and assessed their sex-specific differences in the risk of obesity and a neurocognitive trait.

Here, we aimed to: (1) combine multiple exposure levels to define an environment with high sexual dimorphism in obesity risk; (2) given the correlation between obesity and neurodevelopmental delay in children, we also enquired if the subpopulation exposed to this environment shows a significant sexual dimorphism in neurodevelopment; (3) we then

hypothesized that the individuals who belong to such an environment can be characterized by specific patterns of DNA methylation patterns and aimed to use these patterns to infer subpopulations of high sexual dimorphism in obesity and academic achievement in a large independent cohort of adults.

Methods

Study population

We analyzed data from The Human Early Life Exposome (HELIX). This is a multi-center study that included a total of 1,301 mother-child pairs from six existing birth cohorts in Europe: BIB (Born in Bradford; the United Kingdom) [13], EDEN (Etude des Déterminants pré et postnatals du développement et de la santé de l'Enfant; France) [14], INMA-SAB (Infancia y Medio Ambiente; Spain; subcohort Sabadell) [15], KANC (Kaunas cohort; Lithuania) [16], MoBa (The Norwegian Mother, Father and Child Cohort study; Norway) [17],), and Rhea (Greece) [18] The pairs participated in a common, completely harmonized, follow-up examination, when children were between 5-11 years old [19]. A blood sample was collected, and high dimensional exposure and molecular data were then assessed. In our analyses, we selected the individuals who had data on prenatal exposures, performed the clinical and neurodevelopment examination, and had methylation data (n=1044). All studies received approval from the ethics committees of the centers involved and written informed consent was obtained from all participants. Cohort characteristics are shown in **Table 1**.

Table 1 | Characteristics of the HELIX cohort. Clinical characteristics of children during pregnancy and follow-up.

Children assessed at follow-up	**N=1044**
Sex, male	571 (54.6%)
Age in years, mean (range)	7.9 (5.4-11.9)
Sub-cohort	
BIB (UK)	90 (8.6%)
EDEN (France)	135 (12.9%)
INMA (Spain)	198 (19.0%)
KANK (Lithuania)	196 (18.8%)
MOBA (Norway)	239 (22.9%)
RHEA (Greece)	136 (17.8%)
BMI (kg/m^2), median (range)	16.3 (12.2-29.5)
Obesity	62 (5.9%) - F: 23 (4.9%) - M: 39 (6.8%)
Raven's matrices, median (range)	27 (9-36)
Affected	189 (18.2%) - F: 78 (16.6%) - M: 111 (19.5%)
N-back (2-back accuracy), median (range)	0.91 (0.36-1)
Affected	104 (12.9%) - F: 46 (12.6%) – M: 58 (13.0%)
ANT (accuracy), median (range)	0.97 (0.51-1)
Affected	206 (20.0%) - F: 72 (15.6%) - M: 134 (23.7%)
ADHD	104 (10.0%) - F: 27 (5.8%) - M: 77 (13.6%)
Prenatal characteristics	**N=1044**
Mother's age at pregnancy, mean (range)	30.9 (16-34)
Mother's BMI at pregnancy, median (range)	23.6 (15.8- 51.4)
Maternal education	
Primary school	119 (11.3%)
Secondary school	359 (34.3%)
University degree or higher	566 (54.2%)
Gestational age, median (range)	40 (30.8-44.1)

BMI: Body Mass Index. ANT: Attention Network Test. ADHD: Attention Deficit Hyperactive Disorder. F: Female. M: Male.

Clinical outcomes

Height and weight measurements were converted to body mass index (BMI in kg/m²) for age-and-sex z-scores using the international WHO reference curves to allow comparison with other studies [20]. Obese children were defined as those above the age-and-sex-specific 95th percentile, as recommended by WHO.

Neurodevelopmental outcomes were assessed through a battery of internationally standardized, non-linguistic, and culturally blind computer tests. We assessed working memory, attention and general non-verbal intelligence with the N-back test [21], the attention network test (ANT) [22], and Raven's colored progressive matrices [23]; respectively. The tests were administered in a standardized way by trained field workers through study-provided laptops. The outcomes did not distribute normally. We dichotomized them, taking as cases individuals with outcomes below the first quintiles (20%). We thus studied as clinical outcomes the events of having these cognitive abilities affected. We also considered ADHD diagnosis.

Pregnancy Exposome

HELIX has collected a wide range of environmental contaminant exposures, as well as indicators of the built environment, natural spaces, lifestyle, and noise. Using residential address histories, exposure estimates were assigned for ambient air pollutants, road traffic noise levels, surrounding green and blue spaces, built environment, ultraviolet (UV) radiation, and meteorological variables during pregnancy [19, 24]. Biomarkers of contaminant exposure, including cotinine levels, were measured in appropriate biological samples collected from mothers during pregnancy.

In our study, we included 93 exposures assessed during pregnancy distributed across 17 exposure families. Air pollution, natural spaces, and built environment quantification were assessed during pregnancy by environmental geographic information systems (GIS). Tobacco smoke and diet were evaluated by questionnaires. Exposure variables with their corresponding transformation are described in **Table S1**.

Missing values for all exposures were imputed using the method of chained equations using the *mice* package in R [25], as described in detail elsewhere [26]. When possible, multiple imputation procedure was applied (missing values are imputed stochastically several times). For the imputation process, continuous variables should have a normal distribution. Thus, skewed exposure variables were transformed to achieve normality or categorized if no transformation worked. Exposures with more than 70% of missing values were excluded. Therefore, missing values ranged from 1.5% in traffic density to 65% in fast-food intake during pregnancy. Although none of the participants had complete data on all exposures, 95% of individuals had missing values in less than 30% of exposures.

DNA methylation

One of the main goals of HELIX was to associate multiple environmental factors with omics biomarkers and child health outcomes. For these same children, multi-omics molecular phenotyping was performed, which included measurement of blood DNA methylation (450K, Illumina), among others.

The DNA was obtained from buffy coat collected in EDTA tubes at 5-11 years of age. Briefly, DNA was extracted using the Chemagen kit (Perkin Elmer) in batches of 12 samples. Samples were extracted by cohort and following their position in the original boxes. DNA concentration was

determined in a NanoDrop 1000 UV-Vis Spectrophotometer (ThermoScientific) and with Quant-iTTM PicoGreen® dsDNA Assay Kit (Life technologies). DNA methylation was assessed using the Infinium Human Methylation 450 beadchip (Illumina), following the manufacturer's protocol. Preprocessing of methylation data has been described elsewhere [27]. After sample and probe quality control measures, the number of CpG probes analyzed was 371,533, initially available for 1,192 subjects. We used the Combat algorithm to remove the batch effects supported by the slide. Methylation levels were expressed as beta values and CpG sites were annotated to genes by Illumina HM450 manifest file (version 1.2). We discarded the subjects without exposome data and without European ancestry based on genomic data, resulting in 993 individuals for the methylome analysis. We computed blood cell type proportions following Houseman et al. algorithm [28] and Reinius reference panel [29].

Statistical Methods

Figure 1 shows the statistical workflow.

Identification of prenatal exposures with sexual dimorphism in obesity risk

We used exposome-wide interaction analyses to determine the exposures whose association with obesity was significantly different between sexes. We assessed the associations between obesity (cases and controls) and the interactions between sex (S) and each of the prenatal exposures (D_i) using the logistic regression model

$$E(Y) = logit^{-1}(\alpha_i + \beta_i (S \times D_i) + \Sigma_{r=1\ldots k} \gamma_{ir} C_{ri})$$

where Y is the obesity status of an individual with sex S and i-th exposure D_i. γ_{ir} are the regression coefficients of the k covariates C_{ri} that included sex, exposure I, cohort, year of birth, mother's BMI, mother's weight gain

during pregnancy, gestational age, mother's age at pregnancy, mother's education, whether parents were native from the country cohort, parity, and children age at clinical assessment. B_i were the effects of interest that measure the association between obesity and the interaction between sex and each exposure *i*.

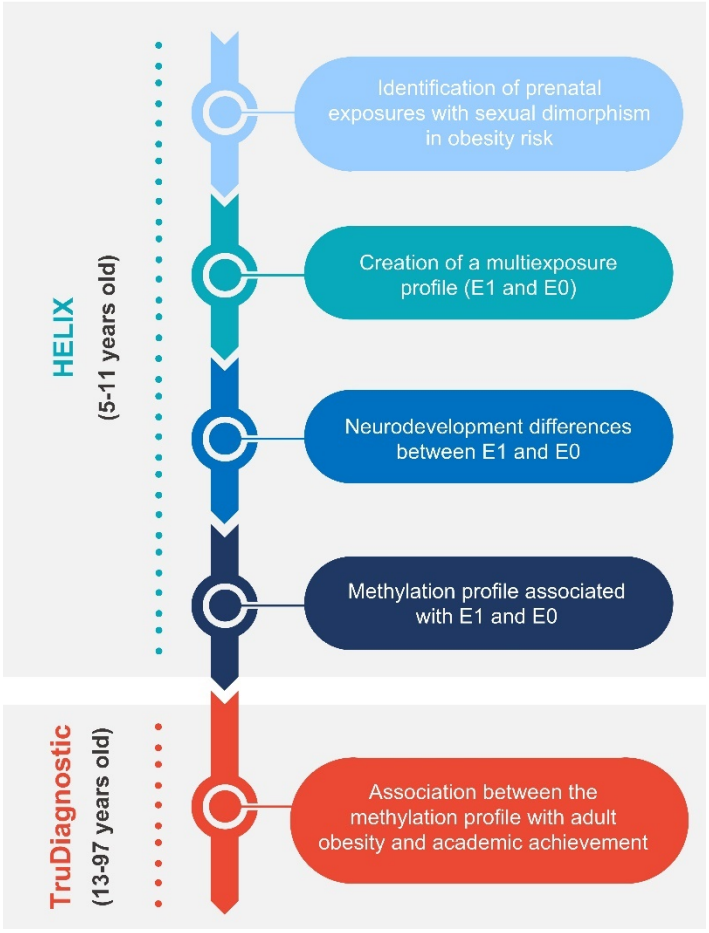


Figure 1 | Statistical workflow. The figure shows all the statistical analyses carried out along the paper and the datasets used for each analysis.

Creation of a multiexposure profile (E1 and E0)

The pregnancy exposures with nominal significant interactions with sex were adjusted by all covariates. Then, we used the residuals as covariates in causal inference modeling, using causal random forest, for the effect of sex on obesity. We aimed to classify individuals into two environments: E1 and E0. The first one (E1) consists of a specific combination of exposure levels that protects girls against obesity. The second one (E0) consists of the remaining combinations of exposure levels. We applied the algorithm *teff*, taking sex as the treatment variable, to define the multiexposure profile associated with E1 [9] (https://teff-package.github.io/).

Neurodevelopment differences between E1 and E0

We used the classification of individuals into E1 and E0 to assess their relationship with sex differences in neurodevelopment. For this analysis, we used logistic regression models on the clinical outcomes (working memory, attention and general non-verbal intelligence with the N-back test, ANT, Raven's colored progressive matrices, and ADHD) and we adjusted by the same covariates as in the epigenome-wide interaction analysis.

Methylation profile associated with E1 and E0

We performed an epigenome-wide association study (EWAS) in the HELIX cohort between E1 and E0. As previously, we used logistic regression models and adjusted all the analyses by the same covariates, adding in this case surrogate variables and counts of different immune cells in the blood. For the latter analyses, we used Bioconductor packages (V.3156) *SVA* and *limma*, and *clusterProfiler* for enrichment analyses. The commented analysis code is available in **Supplementary Methods**.

In order to predict prenatal environments E1 and E0 in independent studies using methylomic data, we used the K-nearest neighbor algorithm from the caret R-library. To build the predictor, we analyzed normalized beta signals but uncorrected by batch effects for unit consistency between studies (beta values). We trained a model in a random selection of 75% individuals from the HELIX data with CpG sites whose associations from the EWAS between E1 and E0 were significant at $P<0.001$ level. We tested the sensitivity and specificity of the predictor in 25% of test samples.

Association between the methylation profile with adult obesity and academic achievement

We aimed to predict E1 and E0 in an independent adult study. As such, we analyzed methylation data from TruDiagnostic DNA biobank that included 3,590 individuals recruited between October 2020 and February 2022. Those individuals have chosen TruDiagnostic for DNA methylation analysis and clinical data has been collected from them. After the quality control, 3,425 participants were included in our study. However, we discarded 166 participants who had missing in at least one CpG site from the methylomic predictor, leading to 3,259 participants. The TruDiagnostic DNA biobank is an EEUU population-based cohort aged between 13 and 97 years old. Among them, 58.7% are male. Since they are adults, obesity in this cohort is defined as a BMI equal to or higher than 30. Cohort characteristics are shown in **Table 2**.

For the DNA methylation analysis, DNA was extracted from peripheral whole blood. The Infinium HumanMethylationEPIC BeadChip was used for DNA methylation assessment following the manufacturer's protocol. Several quality controls and functional normalization were performed using the *meffil* package, resulting in 745,150 probes, as described in detail elsewhere [30]. CpG sites were annotated to genes using EPIC Illumina

annotation ilm 10b4.hg19. Blood cell types were estimated using the blood gse35069 reference panel from the *meffil* package.

Table 2 | Characteristics of the TruDiagnostic DNA Biobank. Clinical outcomes of adults in a large cohort who chose TruDiagnostic for methylomic analysis.

Adult population in TruDiagnostic DNA Biobank	**N= 3259**
Sex, male	1924 (59.0%)
Age in years, mean (range)	52.9 (13.2- 97.8)
BMI (kg/m^2), median (range)	24.9 (10.10- 71.01)
Obesity	446 (13.6%) - F: 120 (3.7%) - M: 326 (10.0%)
Education	
Did not complete high school	30 (0.92%)
High school or equivalent	207 (6.35%)
Technical or occupational certificate	76 (2.33%)
Associate degree	119 (3.65%)
Some college coursework completed	268 (8.22%)
Bachelor's degree	1160 (35.60%)
Master's degree	701 (21.50%)
Doctorate (PhD)	144 (4.41%)
Professional (MD, DO, DDS, JD)	554 (17.0%)
Educational Achievement	2559 (78.5%) - F: 1038 (31.9%) - M: 1521 (46.7%)

BMI: Body Mass Index. F: Female. M: Male. Educational Achievement: completed university degree or higher.

We classified the individuals into E1 and E0 in the TruDiagnostic dataset according to their methylation data based on the predictor trained in HELIX. Although the predictor considered all the CpG sites differentially methylated when comparing E1 and E0, we trained again the predictor using the CpG sites that were common between HELIX and TruDiagnostic. We tested the interaction between sex and the environment on adult obesity

and educational achievement, defined as a completed university degree or higher.

Results

Sexual dimorphism of clinical outcomes

We first assessed whether obesity and the categorized neuropsychological measures were associated with differences between sexes (**Figure 2A**). We fitted logistic regression models adjusting by covariates. Girls showed a lower frequency of obesity than boys, but it was not statistically significant (OR=0.64, P=0.13, see **Figure 2B**). For the neuropsychological measures, we observed that ADHD was lower in girls than boys, consistent with girls' higher protection in attention difficulty. Both associations were statistically significant (OR=0.37, $P=2.87\times10^{-5}$, OR=0.54, $P=4.32\times10^{-4}$). For Raven's matrices and N-back, we did not see significant associations with sex (OR=0.72, P=0.10, OR=0.94, P=0.78) (**Figure 2B**).

Exposome-wide analysis of sex-exposure interactions on obesity

We searched for prenatal exposures that could modulate the association between sex and obesity in childhood. Particularly, we searched for maternal exposure levels in which one sex would be more obese than the other at 5-11 years of age. We performed logistic regressions on obesity for all 93 sex-prenatal exposures interactions, adjusting for all covariates (**Figure 2C-D**). We did not observe any interaction that passed multiple comparison corrections. However, at the nominal level (P<0.05), we observed four interactions between sex (males as reference) and prenatal exposures. First, dairy consumption ($OR_{intreraction}$=2.44, P=0.008), defined as mother's dairy consumption during pregnancy times per week and

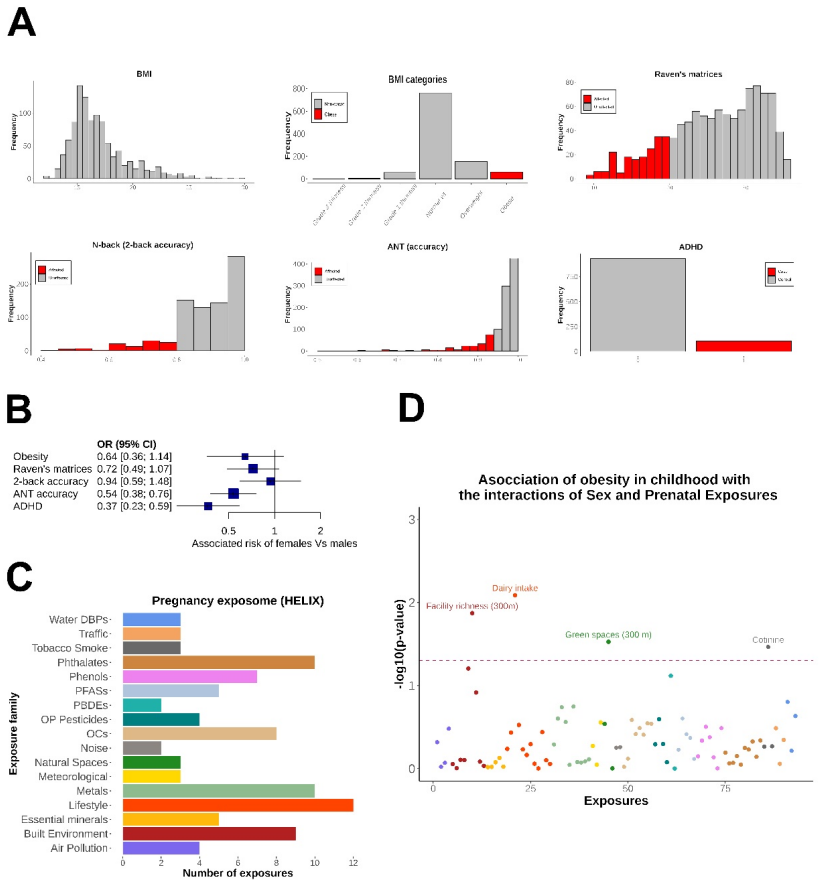


Figure 2 | A) Distributions of clinical outcomes in the HELIX study. Analyses were performed for categorized variables shown in grey (reference) and red. **B) Association of sex with the clinical outcomes, adjusting by covariates**. **C) Number of prenatal exposures in each family measured in HELIX**. **D) Exposome-wide Manhattan plot**. Association of obesity with 93 sex-prenatal exposure interactions (the color follows the exposure family from panel B). The dotted line marks nominal significance (*P*=0.05).

categorized as less than 17.1, between 17.1 and 27.1, and more than 27.1. Second, cotinine levels in mother during pregnancy ($OR_{intreraction}$=1.92, *P*=0.034), classified into three categories: non-smokers (less than 18.32), second-hand smokers (between 18.4 and 48.4), and smokers (more than

50). Third, facility richness ($OR_{intreraction}$=1.11, *P*=0.013), defined as the percentage of different facility types present compared to the maximum potential number of facility types at a 300m buffer during the pregnancy period. Fourth, the presence of green spaces ($OR_{intreraction}$=0.27, *P*=0.029), answering the question of whether the mother lived within a distance of 300m of green space during the pregnancy period. A stratified analysis by sex of the association between obesity and the significant exposures revealed that dairy consumption and cotinine levels were risk factors only for girls (OR= 2.88, *P*= 0.0009; OR=1.91, *P*=0.0128) while facility richness and green spaces were protective and risk factors for boys, respectively (OR= 0.92, *P*= 0.005; OR=5.06, *P*=0.007), see **Figure 3**.

Exposure environment of high differences in obesity risk between sexes

We asked whether a combination of the four significant exposures and their levels could define specific environments where one sex is likely more obese than the other one. The exposure residuals, adjusted by covariates, were used in causal inference modeling, with the aim to classify individuals into environments of high sexual dimorphism in obesity. We considered the multiexposure profile defined by the mother's dairy intake, cotinine levels, living richness facilities, and green spaces during pregnancy. We randomly selected a set of 208 individuals from the HELIX cohort to infer their expected sex-difference in obesity risk given their personal multiexposure profiles. We thus applied the causal modeling algorithm *teff*, taking sex as the treatment variable, and observed 27 children (13 females, 14 males) living in personal environments where girls are less likely obese than boys. By contrast, we found only one boy living in a personal environment where girls are more likely obese than boys (**Figure 4**).

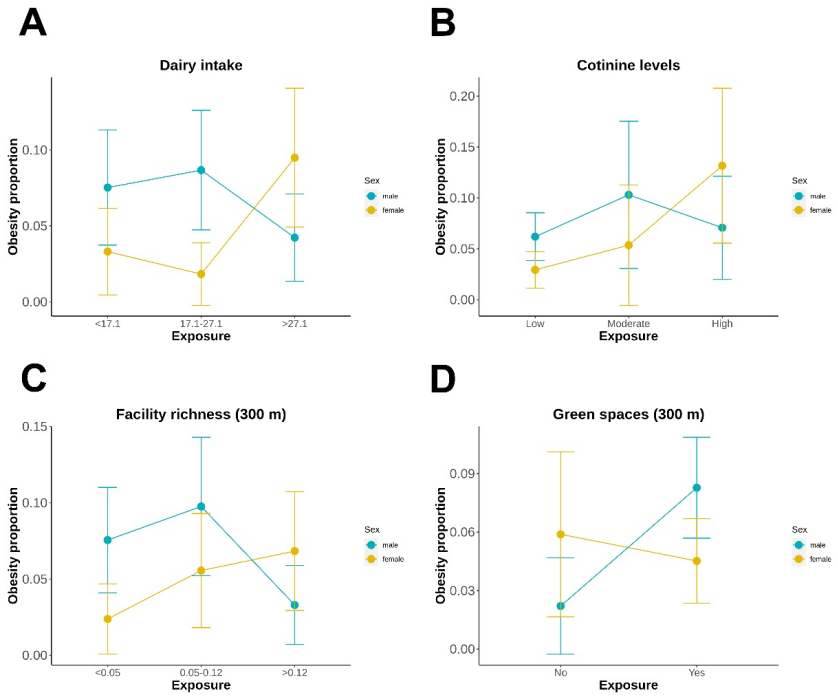


Figure 3 | Sex-exposure interaction plots on obesity. A) Mother's dairy consumption during pregnancy. The figure shows the highest proportion of obesity in girls with the highest level of dairy consumption. **B) Mother's cotinine levels in the blood**. The highest levels of obesity were observed in girls with high cotinine levels. **C) Facility richness in living neighborhoods of pregnant mothers.** A high abundance of facility richness is correlated with a low prevalence of obesity in boys. **D) Green spaces at 300m from pregnant mothers' homes**. The highest prevalence of obesity was observed in boys with mothers living in the presence of green spaces. The bars represent the 95% confidence intervals for the estimated proportion of obesity.

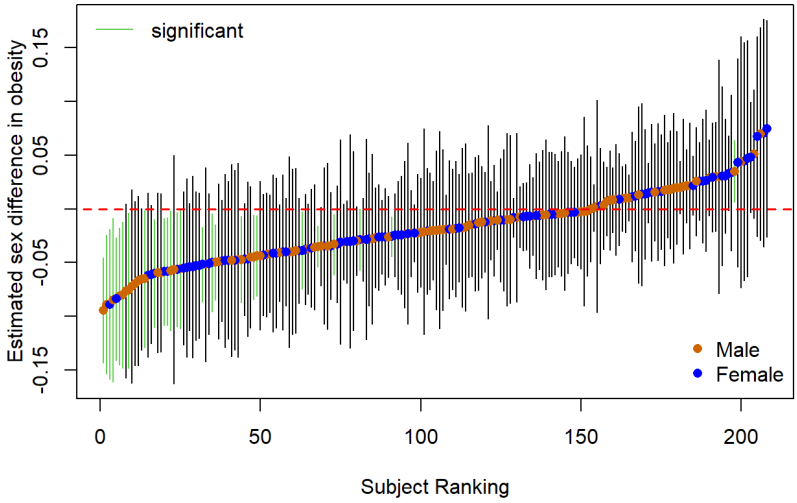


Figure 4 | The estimated sex difference on obesity risk in personal prenatal environments. The personal prenatal exposure environments were defined by the mother's dairy intake, cotinine levels, living richness facilities, and green spaces. The sex of the individual living in a particular prenatal environment is shown in blue (male) and orange (female). The bars show the 95%CI for the effect of a personal prenatal environment on females in relation to males. The intervals were estimated using causal modeling implemented in *teff*. Green lines are significant sex differences in obesity risk given by the prenatal environments.

We then aimed to determine whether the personal environments where girls are less likely obese than boys could be averaged into a single prenatal environment, whose female protection against obesity was stronger than those observed for the individual exposures. We created an average environment with highly significant female protection against obesity, which hereinafter we will refer to it as E1. This environment was obtained using the personal environments of the 27 children where girls are expected to be less obese than boys. E1 was defined as a binary vector, with one entry for each level of the four exposures, indicating whether a given exposure averaged across 27 individuals was higher or lower than the average across

the entire training set of 208 children. We used the multiexposure profile of each child to classify all the individuals in the HELIX cohort depending on whether they belong to the E1 or not. To this end, we used soft targeting that tested whether they matched the environment in at least 60% of the exposures. We observed a total of 675 (64%) individuals classified into E1. All the individuals that did not match E1 were classified into environment E0. E1 was characterized by moderate dairy consumption, low cotinine levels, low abundance of facility richness, and the presence of green spaces (**Figure 5A-D**). Therefore, the environment captured both obesity protection for girls and obesity risk for boys, as expected from the individual exposures.

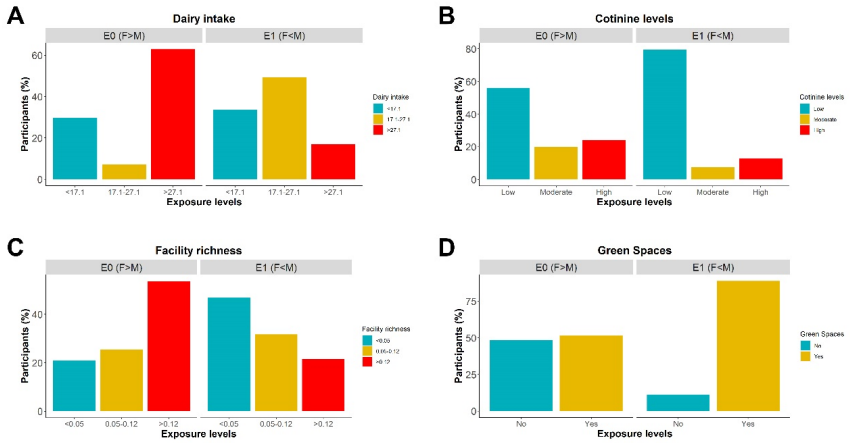


Figure 5 | Characterization of the common prenatal environment where girls are more protected than boys against obesity (E1) against the reference environment (E0). Environment E1 is the common environment of individuals with personal environments where girls are significantly less obese than boys, these are the individuals with green confidence intervals in Figure 4. E1 is defined by low mother dairy intake, low cotinine levels, low richness facilities, and the presence of green spaces. An individual belongs to E0 if he/she does not belong to E1.

We then observed a strong association of the sex-environment interaction on child obesity, adjusting by covariates ($OR_{interaction}$=0.070, P=2.59×10^{-5}). Stratified associations by sex between the environment and obesity risk were also significant (girls: OR= 0.18, P= 4.73×10^{-4}; boys: OR=3.14, P=0.012), suggesting stronger environment gains in the protection for girls than in the risk for boys (**Figure 6A**). These results show that E1 can be regarded as a prenatal environment of female protection against childhood obesity, with much stronger protection than those given its individual exposure components.

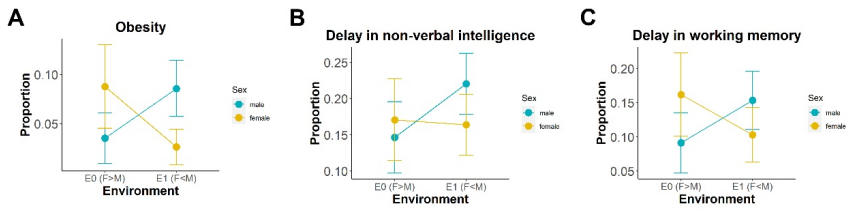


Figure 6 | A) Sex-environment interaction plot on obesity risk. The figure shows that E1 defines a prenatal environment of strong female protection against childhood obesity, across HELIX. **B) Sex-environment interaction plot on raven's matrices underperformance.** Affected individuals are those with outcomes below the first quintiles. **C) Sex-environment interaction plot on N-back underperformance.** Affected individuals are those with outcomes below the first quintiles.

Sexual dimorphism in obesity and neurodevelopment

We asked whether the environment of high differences in obesity between sexes was also an environment of high differences in neurodevelopment. First, we assessed the association between obesity and four neuropsychological outcomes, fitting logistic regression models on obesity and adjusting by all covariates, including sex. We observed that low values

of Raven's matrices and N-back test tests were significant risk factors for obesity (OR= 2.42, *P*=0.01; OR= 2.65, *P*=0.02, see **Figure 6B**), as ADHD diagnosis increased the risk (OR= 2.15, *P*=0.03, see **Figure 6C**). However, we did not find significant associations between obesity and attention outcome.

We tested whether the subject classification into the environments E1 and E0 significantly interacted with sex on each of the neuropsychological outcomes, as it did with obesity. We found that the sex-environment interaction was associated with higher outcomes of both Raven's matrices ($OR_{interaction}$=0.42, *P*=0.047) and N-back test ($OR_{interaction}$=0.31, *P*=0.02), suggesting a higher performance of girls with respect to boys in these two tests, within E1. Associations were fully adjusted by covariates.

Methylomic profile associated with the prenatal environment of high sex-differences in obesity

We aimed to investigate whether the methylome captured the differences between individuals belonging to E1 or E0. We performed an epigenome-wide association study (EWAS) of the classification of children in the prenatal environment, adjusting by covariates and immune cell counts. Methylation data was extracted from blood samples and were previously normalized and corrected for surrogate variation. We did not observe any significant association at a genome-wide level, after correcting for multiple comparisons, see top associations in **Table S2**. We also performed an enrichment analysis for the top associations (nominal *P*<0.01). We tested different GO terms from molecular function, cellular components, and biological processes (**Figure 7**), and observed several pathways related to neuronal processes. Most remarkably, synapse organization (*P*-adjusted=0.0001) and regulation of synapse structure or activity (*P*-

adjusted=0.006) are two biological processes directly related to neurodevelopment.

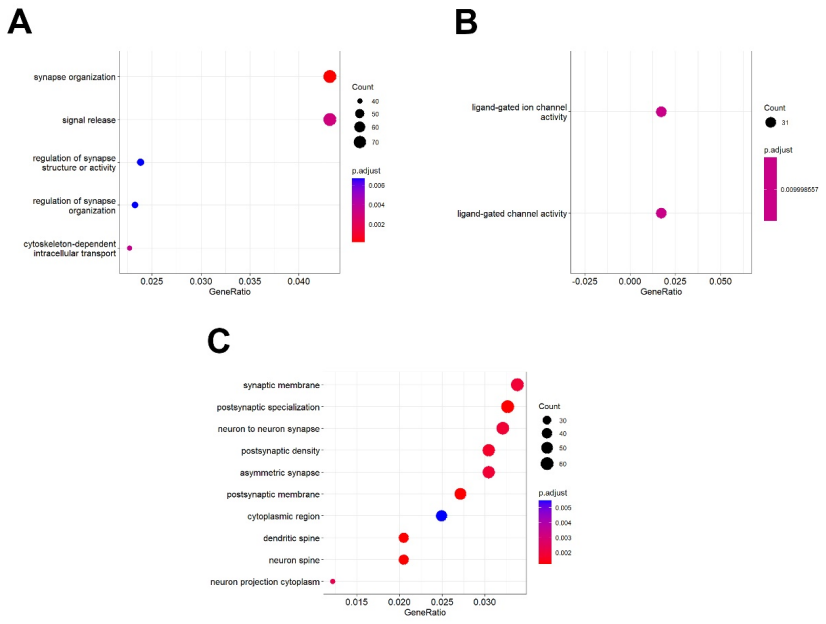


Figure 7 | Enrichment analysis for the differentially methylated sites associated with E1. A) Gene Ontology. B) Cellular components. C) Biological processes. Epigenome-wide analysis for the prenatal environment E1 was performed and methylation probes with associations at P<0.001 were selected. Probes were mapped to genes that were used in enrichment analyses. The analyses mainly show pathways related to neuronal function.

Sex differences in adult obesity and academic achievement

We used methylation data from TruDiagnostic DNA Biobank to determine if adult individuals could be classified into groups whose methylation profiles were consistent with the two prenatal environments in children. We aimed to test whether these methylation profiles were associated with sex differences in obesity and academic achievement in adulthood.

TruDiagnostic data comprises 3,259 adult individuals with methylome data and clinical information, see **Table 2**. We used the K-nearest neighbor algorithm to build a predictor of the environment using the methylation data from HELIX. We selected the CpG sites whose associations with the environment were significant at a nominal *P*-value lower than 0.001 and that were common with TruDiagnostic methylation data (143 CpG sites). We observed a 77% specificity of the predictor on a 25% test set of HELIX individuals, randomly selected to train the predictor.

We used the methylomic predictor on TruDiagnostic data and observed that 1,764 individuals (54.0%) were classified into the group with consistent methylation to the prenatal environment E1. We then tested whether the individuals classified into this group had high differences in obesity when comparing men and women, but we did not observe any significant association (OR=1.14, *P*=0.56). We adjusted by sex, group classification, ethnicity, and age. We also tested the association of the interaction with academic achievement, considering individuals that completed a college degree (78.5%). We found a significant association between the sex-by-group interaction and education achievement (OR=1.59, *P*=0.008, see **Figure 8**), suggesting higher academic achievement of women with respect to men, within this methylation-defined group.

Discussion

We have shown in the HELIX cohort that environments defined by a multiexposure profile with different effects on obesity for each sex can be identified with the novel use of causal inference [9]. In a previous study on the same cohort, no significant associations were observed for individual prenatal exposures with overweight and obesity status, while cotinine levels were associated with BMI only at nominal significance [31]. Although we

observed only four nominally significant interactions between prenatal exposures and sex on obesity, we revealed a prenatal environment defined by specific levels of these exposures whose effect on obesity strongly changed between sexes, with a 93% reduction in obesity risk for girls in relation to boys ($OR_{interaction}$=0.070, P=2.59×10^{-5}). In the environment defined by moderate dairy consumption, low cotinine levels, low facility richness, and the presence of green spaces, girls are more protected than boys against obesity.

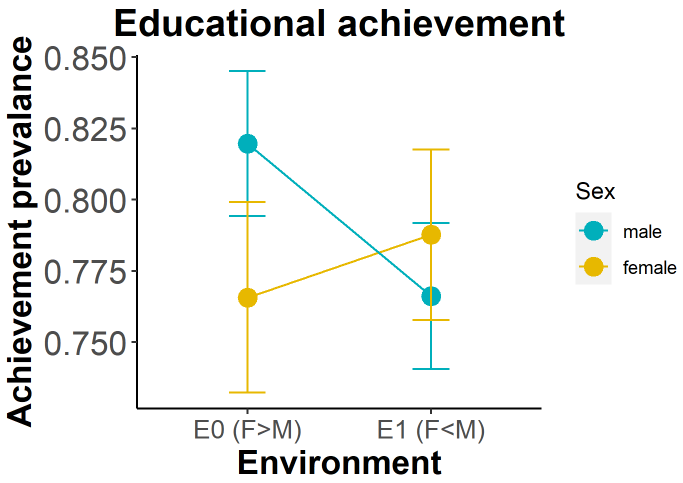


Figure 8 | Sex-methylation profile interaction plot on adult academic achievement in the TruDiagnostic cohort. Individuals were classified into two methylation groups that correlated with two different prenatal environments, being one of them an environment where girls were protected against obesity in the HELIX study. A positive academic achievement was considered for individuals with a university degree.

Previous studies have shown conflictive findings on dairy intake during pregnancy and its relation to long-term body composition of children. Voerman et al. reported significant associations with abdominal fat in

children and strong interaction with sex on the pericardial fat mass index, with a higher risk for girls [32]. However, other studies have reported no significant associations [33, 34]. Our findings suggest that part of the discrepancy could be due to the interaction with sex.

Concerning obesity and cotinine levels in the blood of pregnant mothers, previous studies have shown a 50% increase in childhood overweight for smoking during pregnancy [35], with a dose-response relationship [36]. Cotinine levels have also been associated with low birth weight but rapid gains in BMI after delivery [37]. In a Japanese population, Susuki et al. observed that boys of mothers who smoked during pregnancy had higher gains in BMI trajectories compared with girls [38]. We found, however, higher obesity frequency for girls of mothers with high cotinine levels. In a large study of ~90,000 mother-children pairs, also in Japan [37], they observed that rapid gains in BMI of children were associated with urinary cotinine concentration of mothers but not with self-reported smoking status. While their results were not stratified by sex, it shows that cotinine is a more accurate assessment of pregnancy smoking.

In relation to green spaces, systematic reviews have shown weak evidence for its relationship with children's obesity [39, 40]. Associations of green spaces during pregnancy and their differential effect on sex have not been previously assessed. We found that prenatal green space is a risk factor for boys' obesity only. A recent study of the HELIX cohort showed significant associations between children's overweight and obese status with the built environment (land use mix) [41]. Children living in built environments in absence of green spaces could be at higher risk of obesity (likely due to its relationship with physical activity). However, we observed that a low abundance of facility richness and the presence of green spaces during pregnancy are risk factors for obesity in boys. Both environmental

conditions of the pregnant mother are consistent with less urbanized environments where adult obesity may be more frequent [42].

In this study, we observed that the combination of the specific levels for the four exposures maximizes the differences in obesity risk between girls and boys. Previous studies have already suggested that better prediction of an outcome can be obtained from the aggregation of multiple environmental factors into risk scores [43, 44] or the use of mixture models [45]. In line with this, we used causal inference for classifying the individuals in two environments (E0 and E1) based on the combination of the four exposures.

After classifying individuals in the two environments, we further investigated whether the individuals belonging to the environment with higher sexual dimorphism in obesity presented also sexual dimorphism in neurodevelopmental delay. Based on previous studies, prenatal factors, such as maternal obesity, have been seen associated with both obesity in children and lower cognitive abilities and ADHD [46, 47]. Animal studies have shown that mice whose mothers were on high-fat diets during pregnancy have alterations in brain methylation of dopaminergic and opioid genes [48, 49]. In addition, the neurodevelopmental delay appears to be more frequent in obese boys [11]. A longitudinal prospective study has shown that working memory and attention performance are reduced by increasing BMI in children [50]. Our study offers additional evidence of this relationship, since the environment that protects girls against obesity also protects girls against neurodevelopmental delay, while having the opposite effect in boys. Furthermore, the environment is associated with methylation probes that are enriched in neurodevelopmental pathways, providing more evidence for this hypothesis.

In relation to the long-term association between obesity and academic achievement, a systematic review of longitudinal studies has shown the

strongest evidence for the negative association between adolescent girls' BMI and math attainment [51]. Our analyses support the notion that cognitive abilities involving executive functions may explain part of the association, as they share a common environment that may protect girls against obesity. Furthermore, we observed that the methylomic profile associated with the sexual dimorphic environment was also associated with higher academic achievement in women. However, we did not observe associations with adult obesity. This can be due to the loss of specificity of the predictor as a marker of sexual dimorphism of obesity in adult life. Therefore, further work is required to validate these findings.

Conclusion

We aimed to advance a novel approach to the study of sexual dimorphism, based on high dimensional exposure data and recent methods of causal inference. The methodological approach can also be used to determine the environmental landscape that promotes sexual dimorphisms in studies with high dimensional exposure data.

In summary, girls in childhood may be protected against obesity if their pregnant mothers had moderate dairy consumption, low cotinine levels, and lived in environments with a low abundance of rich facilities and the presence of green spaces. The environment is also protective against the neurodevelopmental delay of non-verbal intelligence and working memory that may be reflected in their adult academic achievement. While female protection is measured against male risk, female protection outweighs the risk of obesity in boys. Our study motivates further public health efforts to raise public awareness of moderating a high-fat diet, and avoiding smoking and second-hand smoking during pregnancy to protect children against obesity and neurodevelopmental delay.

Supplementary Information

Additional file 1: Table S1. Exposure variables in the HELIX cohort. **Table S2.** CpG sites with a nominal P-value lower than 0.001 in the epigenome-wide association study for the classification of children in the prenatal environment (E0 and E1).

Additional file 2: Supplementary Methods.

Acknowledgements

We are grateful to all participants and researchers who took part in this study.

Author contributions

AC and JRG conceived the study. AC designed the analysis. AC and NC-G performed statistical analyses and wrote the first draft of the manuscript. JRG supervised the study and contributed to reviewing and editing the manuscript. MV coordinates the HELIX project, JU is the data manager, and LM is the scientific coordinator. MB, JW, RSl, MC, JRG, and MV designed the omics study in HELIX. The following authors participated in omics data acquisition and quality control: MB (DNA methylation), AnC (DNA methylation), MJN (exposome), and CT (exposome). JW, RG, MV, RSl, LC, and CT are the PIs of the cohorts. TY, SA, MC, JL, NS, and KG participated in sample and data acquisition. VBD, RS, HW, and TLM generated and coordinated the data, secured ethics approval and consents from the TruDiagnostic Biobank. All authors read and approved the final version of the manuscript.

Funding

The study has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 308333 (HELIX project); and the H2020-EU.3.1.2. - Preventing Disease Programme under grant agreement no 874583 (ATHLETE project).

BiB received core infrastructure funding from the Wellcome Trust (WT101597MA) and a joint grant from the UK Medical Research Council (MRC) and Economic and Social Science Research Council (ESRC) (MR/N024397/1). INMA-SAB data collections were supported by grants from the Instituto de Salud Carlos III, CIBERESP, and the Generalitat de Catalunya-CIRIT. KANC was funded by the grant of the Lithuanian Agency for Science Innovation and Technology (6-04-2014_31V-66). The Norwegian Mother, Father and Child Cohort Study is supported by the Norwegian Ministry of Health and Care Services and the Ministry of Education and Research. The Rhea project was financially supported by European projects (EU FP6-2003-Food-3-NewGeneris, EU FP6. STREP Hiwate, EU FP7 ENV.2007.1.2.2.2. Project No 211250 Escape, EU FP7-2008-ENV-1.2.1.4 Envirogenomarkers, EU FP7-HEALTH-2009- single stage CHICOS, EU FP7 ENV.2008.1.2.1.6. Proposal No 226285 ENRIECO, EU- FP7- HEALTH-2012 Proposal No 308333 HELIX), and the Greek Ministry of Health (Program of Prevention of obesity and neurodevelopmental disorders in preschool children, in Heraklion district, Crete, Greece: 2011-2014; "Rhea Plus": Primary Prevention Program of Environmental Risk Factors for Reproductive Health, and Child Health: 2012-15).

This research has received funding from the Spanish Ministry of Science and Innovation through the "Centro de Excelencia Severo Ochoa 2019-2023 (CEX2018-000806-S) program, and support from the Generalitat de

Catalunya through the CERCA Program. NC and JU are supported by Spanish regional program PERIS (Ref.: SLT017/20/000061 and SLT017/20/000119, respectively), granted by Departament de Salut de la Generalitat de Catalunya. TruDiagnostics also provided funding for data analysis.

Availability of data and materials

Any custom code or software used in our analysis is available at DOI: 10.5281/zenodo.6417926 (URL: https://zenodo.org/badge/latestdoi/296552532).

The HELIX data warehouse has been established as an accessible resource for collaborative research involving researchers external to the project. Access to HELIX data is based on approval by the HELIX Project Executive Committee and by the individual cohorts. Further details on the content of the data warehouse (data catalog) and procedures for external access are described on the project website (http://www.projecthelix.eu/index.php/es/data-inventory). The data used in this analysis are not available for replication because specific approvals from the HELIX Project Executive Committee and the University of Southern California Institutional Review Board must be obtained to access them.

The data from the TruDiagnostic Biobank is available upon reasonable request due to ensure the privacy of the participants. Please e-mail varun@trudiagnostic.com for data requests.

Declarations

Ethics approval and consent to participate

All studies from the HELIX project received approval from the ethics committees of the centers involved and written informed consent was obtained from all participants.

The study from the TruDiagnostic Biobank involving human participants was reviewed and approved by the IRCM IRB. The patients/participants provided their written informed consent to participate in this study.

Consent for publication

Not applicable.

Competing interests

VBD, RS, HW, and TLM are employees of TruDiagnostic, the company that has provided the data from the TruDiagnostic Biobank for this study. JRG has received funding from TruDiagnostic as a scientific advisor. The other authors have nothing to declare.

References

1. Muenchhoff M, Goulder PJR. Sex Differences in Pediatric Infectious Diseases. J Infect Dis. 2014;209 Suppl 3:S120.

2. De Bellis MD, Keshavan MS, Beers SR, Hall J, Frustaci K, Masalehdan A, et al. Sex Differences in Brain Maturation during Childhood and Adolescence. Cereb Cortex. 2001;11:552–7.

3. Shah B, Tombeau Cost K, Fuller A, Birken CS, Anderson LN. Sex and gender differences in childhood obesity: contributing to the research agenda. BMJ Nutr Prev Heal. 2020;3:387–90.

4. Stachenfeld NS, Mazure CM. Precision medicine requires understanding how both sex and gender influence health. Cell. 2022;185:1619–22.

5. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): Project rationale and design. Environmental Health Perspectives. 2014;122:535–44.

6. Zhang P, Carlsten C, Chaleckis R, Hanhineva K, Huang M, Isobe T, et al. Defining the Scope of Exposome Studies and Research Needs from a Multidisciplinary Perspective. Environ Sci Technol Lett. 2021;8:839–52.

7. Mauvais-Jarvis F, Bairey Merz N, Barnes PJ, Brinton RD, Carrero JJ, DeMeo DL, et al. Sex and gender: modifiers of health, disease, and medicine. Lancet (London, England). 2020;396:565.

8. Sadovsky Y, Mesiano S, Burton GJ, Lampl M, Murray JC, Freathy RM, et al. Advancing human health in the decade ahead: pregnancy as a key window for discovery: A Burroughs Wellcome Fund Pregnancy Think Tank. Am J Obstet Gynecol. 2020;223:312–21.

9. Cáceres A, González JR. teff: estimation of Treatment EFFects on transcriptomic data using causal random forest. Bioinformatics. 2022. https://doi.org/10.1093/BIOINFORMATICS/BTAC269.

10. Miller AL, Lee HJ, Lumeng JC. Obesity-associated biomarkers and executive function in children. Pediatr Res. 2015;77:143–7.

11. Wentz E, Björk A, Dahlgren J. Neurodevelopmental disorders are highly over-represented in children with obesity: A cross-sectional study. Obesity (Silver Spring). 2017;25:178–84.

12. Perera F, Herbstman J. Prenatal environmental exposures, epigenetics, and disease. Reprod Toxicol. 2011;31:363–73.

13. Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort profile: The born in bradford multi-ethnic family cohort study. Int J Epidemiol. 2013;42:978–91.

14. Heude B, Forhan A, Slama R, Douhaud L, Bedel S, Saurel-Cubizolles MJ, 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.

15. Guxens M, Ballester F, Espada M, Fernández 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.

16. Grazuleviciene R, Danileviciute A, Dedele A, Vencloviene J, Andrusaityte S, Uždanaviciute I, et al. Surrounding greenness, proximity to city parks and pregnancy outcomes in Kaunas cohort study. Int J Hyg Environ Health. 2015;218:358–65.

17. Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol. 2016;45:382–8.

18. 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;46:1392–3.

19. Maitre L, De Bont J, Casas M, Robinson O, Aasvang GM, Agier L, et al. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. BMJ Open. 2018;8:e021311.

20. De Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J. Development of a WHO growth reference for school-aged children and adolescents. Bull World Health Organ. 2007;85:660–7.

21. Vuontela V, Steenari MR, Carlson S, Koivisto J, Fjällberg M, Aronen ET. Audiospatial and Visuospatial Working Memory in 6–13 Year Old School Children. Learn Mem. 2003;10:74–81.

22. Rueda MR, Fan J, McCandliss BD, Halparin JD, Gruber DB, Lercari LP, et al. Development of attentional networks in childhood. Neuropsychologia. 2004;42:1029–40.

23. Raven JC, Raven J. Progressive matrices couleur/colored progressive matrices. Paris: Centre de Psychologie Appliquée.; 1998.

24. Wild CP. Complementing the Genome with an "'Exposome'": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology. 2005;14:1847–50.

25. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45:1–67.

26. Tamayo-Uria I, Maitre L, Thomsen C, Nieuwenhuijsen MJ, Chatzi L, Siroux V, et al. The early-life exposome: Description and patterns in six European countries. Environ Int. 2019;123:189–200.

27. Carreras-Gallo N, Cáceres A, Balagué-Dobón L, Ruiz-Arenas C, Andrusaityte S, Carracedo Á, et al. The early-life exposome modulates the effect of polymorphic inversions on DNA methylation. Commun Biol 2022 51. 2022;5:1–13.

28. 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 Bioinforma 2012 131. 2012;13:1–16.

29. 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:[e41361].

30. Carreras-Gallo Natàlia. Supplementary Material Carreras-Gallo, 2022: TruDiagnostic methylation data pre-processing. https://zenodo.org/badge/latestdoi/29655253. Accessed 14 Oct 2022.

31. Vrijheid M, Fossati S, Maitre L, Márquez S, Roumeliotaki T, Agier L, et al. Early-life environmental exposures and childhood obesity: An exposome-wide approach. Environ Health Perspect. 2020;128:1–14.

32. Voerman E, Gaillard R, Geurtsen ML, Jaddoe VWV. Maternal First-Trimester Cow-Milk Intake Is Positively Associated with Childhood General and Abdominal Visceral Fat Mass and Lean Mass but Not with Other Cardiometabolic Risk Factors at the Age of 10 Years. J Nutr. 2021;151:1965–75.

33. Hrolfsdottir L, Rytter D, Hammer Bech B, Brink Henriksen T, Danielsen I, Steingrimsdottir L, et al. Maternal milk consumption, birth size and adult height of offspring: a prospective cohort study with 20 years of follow-up. Eur J Clin Nutr. 2013;67:1036–41.

34. Leary SD, Ness A, Emmett P, Smith GD. Maternal diet in pregnancy and offspring height, sitting height, and leg length. J Epidemiol Community Health. 2005;59:467–72.

35. Oken E, Levitan EB, Gillman MW. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. Int J Obes (Lond). 2008;32:201–10.

36. Møller SE, Ajslev TA, Andersen CS, Dalgård C, Sørensen TIA. Risk of childhood overweight after exposure to tobacco smoking in prenatal and early postnatal life. PLoS One. 2014;9.

37. Hirai H, Okamoto S, Masuzaki H, Murata T, Ogata Y, Sato A, et al. Maternal Urinary Cotinine Concentrations During Pregnancy Predict Infant BMI Trajectory After Birth: Analysis of 89617 Mother-Infant Pairs in the Japan Environment and Children's Study. Front Endocrinol (Lausanne). 2022;13.

38. Suzuki K, Kondo N, Sato M, Tanaka T, Ando D, Yamagata Z. Gender differences in the association between maternal smoking during pregnancy and childhood growth trajectories: multilevel analysis. Int J Obes (Lond). 2011;35:53–9.

39. Jia P, Cao X, Yang H, Dai S, He P, Huang G, et al. Green space access in the neighbourhood and childhood obesity. Obes Rev. 2021;22 Suppl 1 Suppl 1.

40. Lachowycz K, Jones AP. Greenspace and obesity: a systematic review of the evidence. Obes Rev. 2011;12.

41. de Bont J, Márquez S, Fernández-Barrés S, Warembourg C, Koch S, Persavento C, et al. Urban environment and obesity and weight-related behaviours in primary school children. Environ Int. 2021;155.

42. Okobi OE, Ajayi OO, Okobi TJ, Anaya IC, Fasehun OO, Diala CS, et al. The Burden of Obesity in the Rural Adult Population of America. Cureus. 2021;13.

43. Padmanabhan JL, Shah JL, Tandon N, Keshavan MS. The "polyenviromic risk score": Aggregating environmental risk factors predicts conversion to psychosis in familial high-risk subjects. Schizophr Res. 2017;181:17–22.

44. Jeon EJ, Kang SH, Piao YH, Kim SW, Kim JJ, Lee BJ, et al. Development of the Korea-Polyenvironmental Risk Score for Psychosis. Psychiatry Investig. 2022;19:197–206.

45. Güil-Oumrait N, Cano-Sancho G, Montazeri P, Stratakis N, Warembourg C, Lopez-Espinosa MJ, et al. Prenatal exposure to mixtures of phthalates and phenols and body mass index and blood pressure in Spanish preadolescents. Environ Int. 2022;169:107527.

46. Edlow AG. Maternal obesity and neurodevelopmental and psychiatric disorders in offspring. Prenat Diagn. 2017;37:95–110.

47. Contu L, Hawkes CA. A Review of the Impact of Maternal Obesity on the Cognitive Function and Mental Health of the Offspring. Int J Mol Sci. 2017;18.

48. Grissom NM, Herdt CT, Desilets J, Lidsky-Everson J, Reyes TM. Dissociable deficits of executive function caused by gestational adversity are linked to specific transcriptional changes in the prefrontal cortex. Neuropsychopharmacology. 2015;40:1353–63.

49. Vucetic Z, Kimmel J, Totoki K, Hollenbeck E, Reyes TM. Maternal high-fat diet alters methylation and gene expression of dopamine and opioid-related genes. Endocrinology. 2010;151:4756–64.

50. Li N, Yolton K, Lanphear BP, Chen A, Kalkwarf HJ, Braun JM. Impact of Early-Life Weight Status on Cognitive Abilities in Children. Obesity (Silver Spring). 2018;26:1088–95.

51. Martin A, Booth JN, McGeown S, Niven A, Sproule J, Saunders DH, et al. Longitudinal Associations Between Childhood Obesity and Academic Achievement: Systematic Review with Focus Group Data. Curr Obes Rep. 2017;6:297–313.

6 | DISCUSSION

6.1. General discussion

In this thesis, we have gone one step further in the integration of genetic and environmental exposures in the study of DNA methylation and its influence on complex diseases. To meet this goal, I required abilities in multiple areas, including genetics, bioinformatics, epidemiology, statistics, and programming, which I have learnt during my Genetics Degree, Bioinformatics Master and during this PhD in Biomedicine. In this last step, I had the opportunity to train all these abilities participating in several courses and interacting with members of the Bioinformatic Research Group in Epidemiology (BRGE). **Table 3** summarizes the main results of the three papers described in this thesis.

Table 3 | Overview of the manuscripts of this thesis along with the main results.

	Cohorts	Modulator	Outcome	Analysis	Results
1	TruDiagnostic DNA Biobank	Tobacco, alcohol, and marijuana consumption	DNA methylation	EWAS	Tobacco and alcohol: large effects. Marijuana: small effects.
		Alcohol-related CpG sites	Hypertension	Mediation	Significant mediation by 8 CpG sites
2	HELIX and HUVH Biobank	Genomic inversion	DNA methylation	EWAS and PCA	Methylation patterns in inversions
		Inversion-exposure interaction	DNA methylation	EWAS	64 significant interactions
3	HELIX and TruDiagnostic DNA Biobank	Environment-sex interaction	Obesity and Neurodeve-lopment	Causal inference and linear regression	Prenatal environment with sexual dimorphism for obesity and neurodevelopment delay

We first evaluated the effect of the three most consumed drugs - tobacco, alcohol, and marijuana - on DNA methylation. We identified large genome-wide effects for tobacco and alcohol and small effects for marijuana consumption. We then explored whether changes in methylation were mediating the association between heavy alcohol use and hypertension, and we identified 8 CpG sites significantly mediating this association. Next, we moved on to investigate the effect of common polymorphic inversions on DNA methylation. We found a clear methylation pattern associated with the status of the inversion. We also identified a different effect of multiple environmental exposures on DNA methylation according to the inversion genotype. We finally demonstrated that a prenatal environment consisting of a multiexposure profile is associated with a different risk of obesity in boys and girls. This environment can be inferred from a methylomic profile and, therefore, can be used to predict obesity risk in childhood.

To accomplish our objectives, we used population-based studies based on three cohorts: Human Early-Life Exposome (HELIX), TruDiagnostic DNA Biobank, and Hospital Universitari Vall d'Hebron (HUVH) Biobank (**Table 3**). While the HELIX cohort comprises children (6-11 y), the TruDiagnostic DNA Biobank consists of an adult cohort (13-97 y), and the HUVH biobank comprises prenatal heart tissues. In these cohorts, DNA methylation was assessed with the standardized arrays 450K and EPIC that test thousands of CpG sites along the genome (450,000 and 850,000 CpG sites, respectively). We performed epigenome-wide association studies (EWAS) to test the association between each of these CpG sites with our traits of interest. We also used

mediation analyses and causal inference methods for evaluating the clinical relevance of some of our findings.

Altogether, this thesis suggests that the association between DNA methylation and complex diseases should be studied along with the environmental and genetic context of each individual. This thesis provides examples of genetic influences, individual and combined environmental exposure effects, and gene-environment interactions on DNA methylation.

6.2. Effect of drug consumption on DNA methylation

We first hypothesized that lifestyle was a potential modulator of DNA methylation along the genome. Among lifestyle factors, we focused on tobacco, alcohol, and marijuana consumption because they are the three most used drugs worldwide and they have negative effects on health. We performed an EWAS independently for each drug and we identified 528 CpG sites differentially methylated according to tobacco smoking, 2,569 according to alcohol consumption, and 195 suggestive associations (nominal *P*-value < $1 \cdot 10^{-4}$) for marijuana consumption. Interestingly, we found a large overlap between the differentially methylated genes by these three unhealthy lifestyle habits. As mentioned previously, these drugs are associated with common diseases. In particular, heavy alcohol consumption is associated with a higher risk of hypertension. In our research, we found that the top alcohol-related methylation sites were mediating the association between alcohol consumption and hypertension.

6.2.1. Previous research

There is a huge number of published studies investigating the effect of smoking on DNA methylation. While the first studies used small panels of genes [169,170], the first epigenome-wide association study on tobacco was from 2012 [171]. From then on, several studies have demonstrated the DNA methylation changes along the genome associated with tobacco, even when the exposure was during pregnancy [121,122,124,125,172,173]. To date, the EWAS catalog, which is the biggest database of EWAS [174], gathers 30 publications regarding

tobacco consumption. Most of these publications agree that the cg05575921 mapped to *AHRR* and the cg21566642 in the 2q37.1 region are the most significantly associated CpG sites with smoking [6,125,172].

In the case of alcohol consumption, several studies demonstrated the large effect of drinking on DNA methylation [126,128,129,175–177]. Many of these studies found the cg06690548 mapped to the *SLC7A11* promoter as the most associated methylation site to alcohol consumption [126,128,129]. In addition, many liver biomarkers were associated with the methylation level of *SLC7A11*, which suggests that this gene may be implicated in gastrointestinal disturbance after alcohol consumption [128].

Compared with the large research on tobacco and alcohol, a few studies have evaluated the changes in DNA methylation after marijuana consumption. The first study was performed in 2015 and evaluated the effect of cannabis parental exposure on nucleus accumbens of rats where they identified more than a thousand differentially methylated regions [178]. In 2020, Osborne et al. carried out the first EWAS on heavy cannabis consumption in a reduced population (48 consumers and 18 controls) [131]. In the first analysis, they compared cannabis and tobacco users to non-smokers and they found 5 differentially methylated sites. Second, they selected only cannabis users (no tobacco) and did not find differentially methylated sites. In another study, Markunas et al. evaluated the effect of smoking marijuana on more than 2 thousand women at risk of developing breast cancer [132]. Only one CpG mapped to the CEMIP 5' region was identified as significant. Nevertheless, using

the top 50 CpG sites from the EWAS, they designed a biomarker for lifetime cannabis use.

6.2.2. Tobacco, alcohol, and marijuana as modulators of DNA methylation

In our research, we replicated previous studies revealing a high impact of smoking on DNA methylation along the genome. From the top 50 associations in the EWAS catalog, we replicated 46. We also identified 332 new CpG sites that were not previously reported in the EWAS catalog. Additionally, the two CpG sites most associated with tobacco in most studies (cg05575921 and cg21566642) were also the top CpG sites in our study.

We also replicated a large number of CpG sites for alcohol consumption. Interestingly, differentially methylated genes were highly associated with autistic disorder, acquired scoliosis, and curvature of the spine. The set of adverse effects associated with alcohol consumption during pregnancy is known as fetal alcohol spectrum disorder (FASD). One of the main FASD symptoms is autistic-like traits, as well as scoliosis and other musculoskeletal anomalies [179–181]. Our data suggest that DNA methylation may be the link between alcohol consumption and FASD. In our study, we identified cg06690548 mapped at the *SLC7A11* promoter as the most alcohol-related methylation site, like many studies [126,128,129]. Besides, we identified new CpG sites that were not previously reported in the EWAS catalog, such as three sites at *PRPF8, CBS*, and *MBNL2* genes that were in the top-ranked differentially methylated probes. Whereas the main function of *PRPF8* and *MBNL2*

is the regulation of pre-mRNA alternative splicing, *CBS* is in charge of synthesizing cystathionine.

Although the EWAS for marijuana consumption did not reveal significant CpG sites at the Bonferroni adjustment, we found 195 CpG sites at a suggestive *P*-value lower than $1 \cdot 10^{-4}$. These CpG sites mapped to genes that were enriched in paranodal junction assembly, myelin assembly, and neuromuscular process controlling balance. This suggests a possible implication of DNA methylation changes on the long-term neurotoxic effects of marijuana smoking.

6.2.3. Drug consumption and hypertension mediation

In TruDiagnostic data, we found a slight association between smoking and hypertension and a strong association between alcohol consumption and hypertension, as previously seen [182–185]. Besides, marijuana did not show a significant association with hypertension, in line with previous research that revealed ambiguous associations [186–188].

While light to moderate drinking was not associated with high blood pressure in our study, heavy drinking significantly increased the risk of the disease, as already seen before [185,189]. We performed a multivariate mediation analysis to see whether DNA methylation was mediating the effect of heavy alcohol consumption on hypertension. We found 8 CpG sites as potential mediators that included the six most associated CpG sites with alcohol use. All these CpG sites had lower methylation in heavy consumers compared with no consumers, in line with the global hypomethylation in hypertensive patients revealed previously [62]. Additionally, the low methylation of two of the

potential mediators (cg06690548 at *SLC7A11* and cg14476101 at *PHGDH)* was associated with systolic and diastolic blood pressure in multiple studies [64–66]. Indeed, hypomethylation of these CpG sites was associated with higher expression of their respective genes [66].

SLC7A11 encodes a protein responsible for enhancing antioxidant defense and protecting against endothelial dysfunction and vascular inflammation. This leads to an increased vascular tone and rigidity, that ultimately leads to high blood pressure. Moreover, Richard et al. demonstrated the association between methylation and expression of *SLC7A11* with blood pressure [65]. In the univariate mediation using cg06690548 as a mediator, we found that the methylation level of this CpG site mediated 73.6% of the effect of heavy alcohol on hypertension (*P*-value = 0.008).

The transcription of *PHGDH* results in an enzyme involved in the early steps of serine synthesis, which is associated with tissue growth. Low methylation levels of this gene may participate in vascular adaptation to body-tissue growth during adolescence [66].

Overall, we demonstrated that hypomethylation of CpG sites that are correlated with alcohol consumption and hypertension may mediate the effect of heavy drinking on hypertension. This finding provides new targets to prevent and manage hypertension in individuals with regular alcohol consumption.

6.3. Effect of polymorphic inversions on DNA methylation

Beyond lifestyle factors, we hypothesized that genetic variants also influenced DNA methylation. Although many studies demonstrated the association between SNPs and DNA methylation changes, there are a few studies on structural variants' impact. We focused on polymorphic inversions since they are common in humans, englobe multiple genes, and have been associated with complex diseases. We wanted to see whether the three most common inversions at 8p23.1, 16p11.2, and 17q21.31 were correlated with differentially methylated sites within and around the inversion region. We found that approximately 10% of the CpG sites within the inverted regions +/- 1Mb were significantly differentially methylated according to the inversion genotype. These differences were reflected in distinctive methylation patterns supported by the inversion status. We also validated the effect of inversions on methylation at prenatal stages in heart tissue. Finally, we suspected that the effect of genomic inversions on DNA methylation may differ between individuals due to different environmental exposures. We found 64 significant interactions involving many exposure families, suggesting the important contribution of genomic inversions to gene-environment interactions.

6.3.1. Previous research

Recent studies have demonstrated that polymorphic inversions are important contributors to the genetic context of complex diseases in humans, such as obesity, diabetes, asthma, and cancer, among others

[90–97]. The three most common inversions in humans include 8p23.1, 16p11.2, and 17q21.31. They are polymorphic, non-recurrent, large, and are associated with multiple diseases, including those co-occurring with obesity [91,94]. Due to their big size, they can encapsulate multiple genes and modulate their expression across multiple tissues [94,190–192]. Nonetheless, it is unknown whether genomic inversions have distinctive methylation patterns in the inversion region. Few previous studies have investigated methylation changes at specific sites when studying the association between an inversion and a disease. For instance, Ruiz-Arenas et al. reported a significant mediation between inv-17q21.31 and colorectal cancer using the methylation level of specific CpG sites as mediators [92]. Another two ideas that support our hypothesis are that Shi et al. found methylation patterns in copy number variants [102] and that Shanta et al. reported a big influence on 3D DNA structure by large structural variants [193].

6.3.2. DNA methylation patterns associated with inversion genotypes

We identified distinctive methylation patterns in blood across the inverted regions for the human polymorphic inversions at 8p23.1, 16p11.2, and 17q21.31. We found around 10% of the CpG sites in the inversion region and surroundings to be differentially methylated in all the inversions. We also identified differentially methylated sites at prenatal stages in heart tissue, suggesting that the impact of inversions is relevant during development even *in utero*.

It is worth mentioning that the individual CpG associations that we identified with the inversions may be due to the inversion itself or due

to local genetic variability in linkage with the inversion. However, we carried out a principal component analysis that revealed a spatial pattern given by the correlation of several CpG site associations that fits the extension of the inversion. In this case, the only possibility is that such an extended pattern is due to the inversion itself, likely as a result of the combination of DNA reconfiguration and the accumulation of specific genetic variability along the segment that results from the suppression of recombination between inversion states. These patterns may be caused by differences in the three-dimensional (3D) DNA configuration for each allele [193]. The different configurations of DNA in some haplotypes may produce modifications in the accessibility of the factors that facilitate DNA methylation. This may explain the association between the recurrent and non-polymorphic inversion at Xq28 causing Hemophilia A with specific methylation changes [98]. Another example is the hypermethylation at *de novo* inversions, such as the inversion at 11p15.5 causing Beckwith-Wiedemann [99].

Although inversions at 8p23.1 and 17q21.31 were strongly characterized by their methylation patterns, inversion at 16p11.2 was associated with a less strong pattern. This can be explained by the higher number of haplotypes supported by inv-16p11.2 (two haplotypes in the standard allele and one in the inverted allele) and by the smaller size compared with the other inversions (0.45Mb versus 0.9Mb for inv-17q21.31 and almost 4Mb for inv-8p23.1) [94].

Overall, we suggest that the association between genomic inversions and common diseases may be mediated by the methylation patterns in the inversion region. Further studies may investigate the possible correlation between inversion haplotypes and 3D configurations.

6.3.3. Early-life exposome modulates the effect of genomic inversions on DNA methylation

We further observed that the methylation patterns due to inversion genotypes were modifiable by environmental exposures. We identified multiple interactions between inversions and environmental exposures affecting DNA methylation across the inverted region. Most of the exposures involved in significant interactions were metals, diet, phenols, and organochlorines.

Among the significant interactions, we highlight three of them due to potential clinical interest and substantial support from previous studies. All of these interactions involve the inversion at 8p23.1.

First, we identified the interaction between inv-8p23.1 and meat intake associated with *TDH* methylation. This interaction was interesting due to the independent association of all the factors (inversion, exposure, and gene) with obesity in adults [91,194–196]. In our data, we found that increased intake of meat was associated with lower methylation levels at two CpG sites only in non-inverted homozygous individuals. Interestingly, individuals with this genotype are more at risk of obesity [91]. We propose to consider the methylation of *TDH* and the inversion genotype of individuals who aim to reduce obesity by managing meat intake.

Second, we observed that non-inverted homozygous individuals presented hypermethylation at cg26020513 within *GATA4* when manganese exposure increased. It is worth noting that hypermethylation of this CpG has been strongly associated with congenital heart defects in fetuses [197], mutations in *GATA4* are correlated with cardiac septal

defects [198], and previous studies have demonstrated heart toxicity by high exposure to manganese [199]. Altogether, this interaction deserves further scrutiny.

Finally, we identified different interactions involving the same inversion and the same gene. We observed that the effects of parental tobacco smoke (during pregnancy or in childhood) and air pollution (outdoor PM2.5 exposure) on *TRMT9B* methylation were different according to the genotype of 8p23.1 inversion. Smoking and air pollution have been largely associated with respiratory diseases [200–202]. Moreover, *TRMT9B* is associated with an upper respiratory tract disease [203,204]. Since inv-8p23.1 and asthma have been seen correlated, our results suggest a likely role of gene-environment interaction in this association.

Although validation of the significant interactions remains to be evaluated, we suggest that carriers of specific genetic variants may be more susceptible to (or protected against) disease or developmental disorders if exposed to a relevant environmental risk factor.

6.4. Prenatal environment with high sexual dimorphism

Obesity is a complex disease whose prevalence differs between boys and girls. Genetic and environmental factors are important contributors to obesity risk. As the main difference between individuals is sex, exposome studies aiming at improving precision medicine cannot do without considering how environmental factors affect sexual dimorphism in disease. Recent research has shown that maternal factors during pregnancy can affect disease outcomes later in life. Therefore, we hypothesized that the aggregation of multiple prenatal exposures and behaviors may promote later sexual dimorphism in obesity. We found that girls are more protected than boys against obesity in a prenatal environment defined by moderate dairy consumption, low cotinine levels, low richness facilities, and the presence of green spaces. We then showed that this environment also protected girls from neurodevelopment delay. Finally, we developed a predictor based on methylation to classify individuals into the protector environment or not. We tested the predictor in an independent adult population, and we found that the environment was associated with a high dimorphism in academic achievement.

6.4.1. Previous research

Previous studies have evaluated the association between dairy intake during pregnancy and obesity during life. However, they showed contradictory results. Voerman et al. revealed a significant association between dairy intake and abdominal fat in children, as well as a strong

interaction between dairy intake and sex on the pericardial fat mass index, being girls more at risk than boys [205]. Other papers have shown no significant association [206,207]. Our results suggest that the interaction with sex may be the clue for obtaining replicative results.

Independent studies have investigated the association between cotinine levels during pregnancy and obesity in the life course. Oken et al. revealed an increment of 50% in children overweight when mothers smoked during pregnancy [208]. Moreover, another study demonstrated that this association is dose-dependent, being more overweighted those children with mothers who smoked more cigarettes [209]. Other studies did not evaluate the effect of smoking but cotinine levels in the blood and found a significant association with low birth weight and rapid gains in BMI after delivery [210]. In Japan, boys had higher gains in BMI trajectories compared with girls when their mothers smoked during pregnancy [211]. In a study with more than 90,000 mother-child pairs from Japan they observed that while cotinine levels in urine during pregnancy were associated with BMI of children, self-reported smoking status was not associated with BMI [210]. It seems, therefore, that cotinine is a more accurate assessment of smoking during pregnancy.

Studies involving green spaces revealed weak evidence for the association with children's obesity [212,213]. Stratified studies by sex for evaluating the effect of green spaces during pregnancy on BMI have not been previously assessed. One study using HELIX data reported a significant association between obesity and the built environment [214]. Although the presence of green spaces may be associated with higher physical activity, less urbanized areas with low facility richness and a

higher abundance of green spaces are characterized by more prevalence of adult obesity [215].

On the other hand, previous studies have reported the association between obesity and neurodevelopmental delay [216]. Maternal obesity is associated with obesity in children, lower cognitive abilities, and ADHD [217,218]. Besides, a longitudinal study revealed that increasing BMI in children can result in a reduction in working memory and attention performance [219].

Finally, a systematic review showed that adolescent girls' BMI is negatively associated with math attainment, suggesting an association between obesity and academic achievement [220].

6.4.2. Environment with sexual dimorphism in obesity and neurodevelopment

We first identified four exposures during pregnancy significantly interacting with sex on obesity: dairy intake, cotinine levels, facility richness, and green spaces. We observed that a multiexposure profile defined by the combination of these exposures at specific levels was more significant than the independent exposures. This environment defined by moderate dairy consumption, low cotinine levels, low richness facilities, and the presence of green spaces was associated with protection against obesity only in girls.

Furthermore, we tested whether the environment presented sexual dimorphism for neurodevelopment traits. We found that girls in this environment had less delay in non-verbal intelligence and working

memory compared with boys, in consistent with previous studies [216–218].

6.4.3. Prediction of the environment based on methylomic data

We aimed to investigate whether the methylome captured the differences between individuals belonging and not belonging to the protector environment for girls. We performed an EWAS based on the classification of children in the prenatal environment, and we did not observe any significant association at the genome-wide level after correcting for multiple comparisons. However, the enrichment analysis of the CpG sites with a nominal *P*-value below 0.01 revealed pathways related to neuronal processes, such as synapse organization and regulation of synapse structure or activity. This suggests that DNA methylation may be mediating the interaction between the environment and sex on neurodevelopment traits.

We then investigated whether DNA methylation profiles could be used to classify individuals in independent cohorts into belonging to the protector environment or not. We trained the algorithm in the HELIX cohort where we had environmental and methylation data, and we then tested the classifier in TruDiagnostic data where only methylation data was available. We used the K-nearest neighbor algorithm to build the predictor based on the CpG sites with a nominal *P*-value below 0.01 and that were common with TruDiagnostic methylation data (143 CpG sites). We observed a 77% specificity of the predictor on a 25% test-set of HELIX individuals, randomly selected to train the predictor. This lower-than-expected accuracy of the predictor was likely due to the methylation levels uncorrected for surrogate variation which were used

to increase translation into TruDiagnostic. It may also be due to the low number of CpG sites overlapped between both datasets.

We applied the classifier to TruDiagnostic, and we found that 54% of the individuals were classified into the environment where girls are less obese than boys. We did not observe differences in obesity comparing men and women belonging to the environment. However, we observed a significant interaction between the environment and sex on academic achievement. Women belonging to the protector environment had a higher academic achievement compared with men in the same environment.

6.4.4. Clinical implications

In summary, girls in childhood may be protected against obesity if their pregnant mothers had moderate dairy consumption, low cotinine levels, and lived in environments with a low abundance of rich facilities and the presence of green spaces. This environment is also protective of non-verbal intelligence and working memory delays, which may be reflected in their adult academic achievement.

Our study motivates further public health efforts to raise awareness of moderating a high dairy diet and avoiding smoking and second-hand smoking during pregnancy to protect children against obesity and neurodevelopment delay, particularly in girls.

6.5. Limitations and strengths

While this thesis is subject to several limitations, it also has notable strengths, as described below.

6.5.1. Limitations

- **DNA methylation was obtained from blood samples in all the cohorts.** Then, further research is required to understand the implication of the identified markers in each tissue.

- **DNA methylation only at a one-time point.** The DNA methylation process is dynamic and changes according to genetic and environmental factors throughout life. Studying only one-time points may ignore other methylation levels that can be present in specific time intervals.

- **Different arrays for assessing DNA methylation.** In our data, while HELIX and HUVH Biobank used the 450K array, TruDiagnostic used the EPIC array. These differences may difficult the application of algorithms tested in one cohort and applied to another cohort.

- **Non-availability of genetic data in TruDiagnostic.** The effect of drug consumption on DNA methylation was evaluated in TruDiagnostic. However, genetics has an important role in substance use predisposition. Further research should be done to remove the differentially methylated probes that are a consequence of genetic differences and not to the exposure itself.

- **Heterogeneity within population studies.** The batch effect is always an uncontrollable factor. While several methods have been applied for removing the batch effect, such as DNA methylation normalization and surrogate variable analyses, there is not a standard method suitable for all the datasets.

- **Self-reported drug consumption.** The drug consumption was self-assessed and not specific for a time period, limiting the credibility. In addition, there was no information on whether marijuana was smoked mixed or not with tobacco. This information could benefit future studies on removing the tobacco effect.

6.5.2. Strengths

- **Big sample size.** All the objectives of this thesis have been studied in large populations. The HELIX project comprises 1,301 mother-child pairs and the TruDiagnostic Biobank includes 3,890 adults. Moreover, the heart tissue from 40 fetuses of terminated pregnancies is especially valuable due to the difficulty of obtaining these samples.

- **Multi-omics datasets.** In addition to big sample sizes, these populations have hundreds of variables for each individual. The HELIX project has information relating to the exposome and the child health outcomes, as well as DNA methylation, gene expression, miRNA expression, plasma proteins, serum metabolites, urinary metabolites, and DNA microarray. Besides, TruDiagnostic collected information for personal information,

medical history, social history, lifestyle, and family history. Although during our study we only had access to DNA methylation data, the company is now collecting information for the metabolome and the genome.

- **Evaluation of clinical implication.** As mentioned in the previous strength, we had information for health outcomes. This allowed us to evaluate the clinical implication of our findings in the same population.

- **Standardized methods to measure the exposome.** HELIX is a big project involving six European countries. They applied the same standardized methods to assess the exposure to different hazards in each cohort. A detailed description of the methods used is described elsewhere [166].

- **The validation of previous studies allows for improvement in the credibility of new findings.** In TruDiagnostic data, we were able to evaluate the effect of different drugs on the same individuals using the same normalized methylation values. Since our results for tobacco and alcohol were comparable with previous studies, we may assume that the new findings for marijuana consumption may be reliable. The same happened with the significant mediation between alcohol and hypertension, which implicated sites that were previously seen associated with alcohol or hypertension.

- **Correction for multiple comparisons.** All the *P*-values in the different studies were adjusted by Bonferroni, following a conservative approach.

6.6. Further research work

DNA methylation is a dynamic process with huge implications for disease emergence. In this thesis, we evaluated different modulators of DNA methylation, including genetic and environmental factors. However, further work needs to be done to establish the clinical implications of these modifications.

First, the evaluation of DNA methylation changes by marijuana consumption should be repeated in a larger consumer population with consumption patterns better defined. We found that the top CpG sites were associated with neurological processes. Then, further work is needed to fully understand the implications of these modifications in the mediation between high marijuana consumption and neurological disorders.

Another application of our investigation is related to the estimation of biological age based on DNA methylation. As it is extensively known, many lifestyle factors accelerate biological age. Further analyses should be performed to develop a predictive model based on multiple methylation sites to determine whether an individual has been exposed to tobacco, alcohol, and/or marijuana. This predictor will better estimate the effect of these drugs on biological mechanisms compared to self-reported questionnaires. Then, the biological age acceleration of heavy consumers may be predicted if they do not stop consuming drugs.

Our research has thrown up the possibility of designing a methylomic biomarker to manage and prevent cardiovascular disease in heavy drinkers. Further research, using a larger group of individuals with hypertension and high alcohol consumption, could shed more light on

this hypothesis. In the next years, in collaboration with TruDiagnostic, we will work on the design of this product that will be commercialized by the company.

We demonstrated that inversions show differentiable methylation patterns along the inverted region. Since DNA methylation is involved in chromatin structure regulation, these methylation patterns may produce different tridimensional DNA structures for each allele. The possible association between inversion alleles and different 3D configurations should be investigated in future studies.

Recent studies have demonstrated that common inversions are associated with common diseases, such as obesity, asthma, and neurological disorders. Further work is required for evaluating the implication of the allele-specific methylation patterns in the risk of these diseases.

As previously mentioned, one of the main limitations in gene-environment interactions is the difficulty in replicating the results and the vulnerability of publication bias. Therefore, further studies need to be carried out in order to validate the significant interactions between genomic inversions and environmental exposures.

In the last manuscript, we combined multiple exposures to estimate the risk of obesity in girls and boys independently. Since we found a better association for the aggregated exposures compared with the independent variables, it is recommended to use this methodology in future research.

We also defined an environment with high sexual dimorphism in obesity and neurodevelopment in children. We created an algorithm to predict

the environment based on DNA methylation. We tested this predictor in an adult population, and we did not find an association between sexual dimorphism in obesity in adults and the environment. This can be due to the loss of sexual dimorphism as growing older, or due to the low specificity of the predictor. Therefore, further work is required to validate these findings.

Finally, we have demonstrated that the influence of genetics and environment on DNA methylation should be evaluated simultaneously since the interaction between these factors occurs often.

7 | CONCLUSIONS

7.1.1. Conclusion 1

- **General conclusion**

We revealed that tobacco, alcohol, and marijuana have large effects on genome-wide DNA methylation and that those effects may partially explain the association with neurodevelopment and cardiovascular diseases.

- **Specific conclusions**

❖ We identified 538 CpG sites differentially methylated according to smoking levels, 2,569 according to alcohol consumption, and 195 suggestive associations for marijuana consumption.

❖ Genes differentially methylated for marijuana use were enriched in neurological processes. The overlapped genes between alcohol and tobacco were enriched in signaling events in the nervous system and neurodevelopment. This suggests that methylation may be mediating the association of substance use with neurotoxic effects.

❖ We identified eight CpG sites as potential mediators in the association between heavy drinking and a higher risk of hypertension. Particularly, the methylation levels of cg06690548 mapped to *SLC7A11* mediated 73.6% of this association.

7.1.2. Conclusion 2

- **General conclusion**

We demonstrated that common polymorphic inversions show allele-specific methylation patterns along the inverted region which can be modulated by multiple environmental exposures.

- **Specific conclusions**

❖ During childhood, we found that around 10% of the CpG sites within the inverted regions +/-1 Mb were differentially methylated. Similar changes were observed in prenatal heart tissue, suggesting their relevant role even *in utero*.

❖ The PCA revealed allele-specific patterns given by the correlation of several CpG sites along the inversion region, particularly in 8p23.1 and 17q21.31 inversions.

❖ We identified 64 significant inversion-exposure interactions, suggesting that DNA methylation changes associated with the inversions were modifiable by numerous environmental exposures.

❖ We observed that non-inverted homozygous individuals for inv-8p23.1, those with a higher risk of obesity, had lower methylation levels of two CpG sites (cg01489256 and cg02601489) within the *TDH* gene as meat intake increased.

7.1.3. Conclusion 3

- **General conclusion**

We identified a multiexposure profile with high sexual dimorphism in obesity and neurodevelopment that was reflected in a DNA methylation profile.

- **Specific conclusions**

❖ We recognized a prenatal environment defined by low dairy consumption, low cotinine levels in blood, low abundance of

facility richness, and the presence of green spaces where girls are more protected from obesity than boys.

- The environment is also protective against the neurodevelopmental delay of non-verbal intelligence and working memory for girls.
- We designed a methylation-based predictor of this environment to assess the risk of obesity and neurodevelopment delay in boys and girls in populations without exposome data.
- We found that in adulthood the environment predicted based on DNA methylation is associated with higher academic achievement for women compared to men.

8 BIBLIOGRAPHY

[1] Moore LD, Le T, Fan G. DNA Methylation and Its Basic Function. Neuropsychopharmacol 2013 381 2012;38:23–38. https://doi.org/10.1038/npp.2012.112.

[2] Moosavi A, Ardekani AM. Role of Epigenetics in Biology and Human Diseases. Iran Biomed J 2016;20:246. https://doi.org/10.22045/IBJ.2016.01.

[3] Zoghbi HY, Beaudet AL. Epigenetics and Human Disease. Cold Spring Harb Perspect Biol 2016;8:1–28. https://doi.org/10.1101/CSHPERSPECT.A019497.

[4] Shukla A, Bunkar N, Kumar R, Bhargava A, Tiwari R, Chaudhury K, et al. Air pollution associated epigenetic modifications: Transgenerational inheritance and underlying molecular mechanisms. Sci Total Environ 2019;656:760–77. https://doi.org/10.1016/J.SCITOTENV.2018.11.381.

[5] Kaur G, Begum R, Thota S, Batra S. A systematic review of smoking-related epigenetic alterations. Arch Toxicol 2019;93:2715–40. https://doi.org/10.1007/S00204-019-02562-Y.

[6] Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: findings from the Avon Longitudinal Study of Parents and Children (ALSPAC). Hum Mol Genet 2015;24:2201–17. https://doi.org/10.1093/HMG/DDU739.

[7] Daxinger L, Whitelaw E. Understanding transgenerational epigenetic inheritance via the gametes in mammals. Nat Rev Genet 2012 133 2012;13:153–62. https://doi.org/10.1038/nrg3188.

[8] Lim JP, Brunet A. Bridging the transgenerational gap with epigenetic memory. Trends Genet 2013;29:176–86. https://doi.org/10.1016/J.TIG.2012.12.008.

[9] Stenz L, Schechter DS, Serpa SR, Paoloni-Giacobino A. Intergenerational Transmission of DNA Methylation Signatures Associated with Early Life Stress. Curr Genomics 2018;19:665. https://doi.org/10.2174/1389202919666171229145656.

[10] Schmitz RJ, Lewis ZA, Goll MG. DNA Methylation: Shared and Divergent Features across Eukaryotes. Trends Genet 2019;35:818–27. https://doi.org/10.1016/J.TIG.2019.07.007.

[11] Greenberg MVC, Bourc'his D. The diverse roles of DNA methylation in mammalian development and disease. Nat Rev Mol

Cell Biol 2019 2010 2019;20:590–607. https://doi.org/10.1038/s41580-019-0159-6.

[12] Kornberg RD. Chromatin structure: a repeating unit of histones and DNA. Science 1974;184:868–71. https://doi.org/10.1126/SCIENCE.184.4139.868.

[13] Kouzarides T. Chromatin modifications and their function. Cell 2007;128:693–705. https://doi.org/10.1016/J.CELL.2007.02.005.

[14] Zhou BR, Bai Y. Chromatin structures condensed by linker histones. Essays Biochem 2019;63:75–87. https://doi.org/10.1042/EBC20180056.

[15] du Preez LL, Patterton HG. Secondary structures of the core histone N-terminal tails: their role in regulating chromatin structure. Subcell Biochem 2013;61:37–55. https://doi.org/10.1007/978-94-007-4525-4_2.

[16] Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res 2011 213 2011;21:381–95. https://doi.org/10.1038/cr.2011.22.

[17] Alaskhar Alhamwe B, Khalaila R, Wolf J, Bülow V, Harb H, Alhamdan F, et al. Histone modifications and their role in epigenetics of atopy and allergic diseases. Allergy, Asthma Clin Immunol 2018 141 2018;14:1–16. https://doi.org/10.1186/S13223-018-0259-4.

[18] Kurdistani SK, Tavazoie S, Grunstein M. Mapping global histone acetylation patterns to gene expression. Cell 2004;117:721–33. https://doi.org/10.1016/J.CELL.2004.05.023/ATTACHMENT/D5D33CE9-9E1C-4331-BE93-D0B29DB1A6DA/MMC8.XLS.

[19] Zhang P, Wu W, Chen Q, Chen M. Non-Coding RNAs and their Integrated Networks. J Integr Bioinform 2019;16. https://doi.org/10.1515/JIB-2019-0027.

[20] Kaikkonen MU, Lam MTY, Glass CK. Editor's Choice: Non-coding RNAs as regulators of gene expression and epigenetics. Cardiovasc Res 2011;90:430. https://doi.org/10.1093/CVR/CVR097.

[21] Esteller M. Non-coding RNAs in human disease. Nat Rev Genet 2011 1212 2011;12:861–74. https://doi.org/10.1038/nrg3074.

[22] Wei JW, Huang K, Yang C, Kang CS. Non-coding RNAs as regulators in epigenetics (Review). Oncol Rep 2017;37:3–9.

https://doi.org/10.3892/OR.2016.5236/HTML.

[23] Lam JKW, Chow MYT, Zhang Y, Leung SWS. siRNA Versus miRNA as Therapeutics for Gene Silencing. Mol Ther - Nucleic Acids 2015;4:e252. https://doi.org/10.1038/MTNA.2015.23.

[24] Bird AP. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 1980;8:1499. https://doi.org/10.1093/NAR/8.7.1499.

[25] Feng S, Cokus SJ, Zhang X, Chen PY, Bostick M, Goll MG, et al. Conservation and divergence of methylation patterning in plants and animals. Proc Natl Acad Sci U S A 2010;107:8689–94. https://doi.org/10.1073/PNAS.1002720107/SUPPL_FILE/PNAS.201002720SI.PDF.

[26] Yong WS, Hsu FM, Chen PY. Profiling genome-wide DNA methylation. Epigenetics Chromatin 2016 91 2016;9:1–16. https://doi.org/10.1186/S13072-016-0075-3.

[27] Lu Q, Qiu X, Hu N, Wen H, Su Y, Richardson BC. Epigenetics, disease, and therapeutic interventions. Ageing Res Rev 2006;5:449–67. https://doi.org/10.1016/J.ARR.2006.07.001.

[28] He Y, Ecker JR. Non-CG Methylation in the Human Genome. Annu Rev Genomics Hum Genet 2015;16:55. https://doi.org/10.1146/ANNUREV-GENOM-090413-025437.

[29] Chen L, MacMillan AM, Chang W, Ezaz-Nikpay K, Verdine GL, Lane WS. Direct Identification of the Active-Site Nucleophile in a DNA (Cytosine-5)-methyltransferase. Biochemistry 1991;30:11018–25. https://doi.org/10.1021/BI00110A002/ASSET/BI00110A002.FP.PNG_V03.

[30] Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 2012;13:484–92. https://doi.org/10.1038/NRG3230.

[31] Okano M, Bell DW, Haber DA, Li E. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. Cell 1999;99:247–57. https://doi.org/10.1016/S0092-8674(00)81656-6.

[32] Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. Annu Rev Biochem 2005;74:481–514. https://doi.org/10.1146/ANNUREV.BIOCHEM.74.010904.153721.

[33] Vertino PM, Yen RW, Gao J, Baylin SB. De novo methylation of CpG island sequences in human fibroblasts overexpressing DNA (cytosine-5-)-methyltransferase. Mol Cell Biol 1996;16:4555–65. https://doi.org/10.1128/MCB.16.8.4555.

[34] Wu X, Zhang Y. TET-mediated active DNA demethylation: mechanism, function and beyond. Nat Rev Genet 2017 189 2017;18:517–34. https://doi.org/10.1038/nrg.2017.33.

[35] abcam. DNA methylation and demethylation n.d. https://www.abcam.com/epigenetics/dna-methylation-and-demethylation#DNA methylation (accessed September 2, 2022).

[36] Oda M, Kumaki Y, Shigeta M, Jakt LM, Matsuoka C, Yamagiwa A, et al. DNA Methylation Restricts Lineage-specific Functions of Transcription Factor Gata4 during Embryonic Stem Cell Differentiation. PLoS Genet 2013;9:1003574. https://doi.org/10.1371/journal.pgen.1003574.

[37] Marsit CJ. Influence of environmental exposure on human epigenetic regulation. J Exp Biol 2015;218:71–9. https://doi.org/10.1242/jeb.106971.

[38] Bollati V, Baccarelli A. Environmental epigenetics. Heredity (Edinb) 2010;105:105–12. https://doi.org/10.1038/hdy.2010.2.

[39] Stein RA. Epigenetics and environmental exposures. J Epidemiol Community Health 2012;66:8–13. https://doi.org/10.1136/jech.2010.130690.

[40] Carreras-Gallo N, Cáceres A, Balagué-Dobón L, Ruiz-Arenas C, Andrusaityte S, Carracedo Á, et al. The early-life exposome modulates the effect of polymorphic inversions on DNA methylation. Commun Biol 2022;5. https://doi.org/10.1038/S42003-022-03380-2.

[41] Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nat 2011 4807378 2011;480:490–5. https://doi.org/10.1038/nature10716.

[42] Marasca F, Bodega B, Orlando V. How Polycomb-Mediated Cell Memory Deals With a Changing Environment: Variations in PcG complexes and proteins assortment convey plasticity to epigenetic regulation as a response to environment. Bioessays 2018;40. https://doi.org/10.1002/BIES.201700137.

[43] Jones PA. The DNA methylation paradox. Trends Genet

1999;15:34–7. https://doi.org/10.1016/S0168-9525(98)01636-9.

[44] Laurent L, Wong E, Li G, Huynh T, Tsirigos A, Ong CT, et al. Dynamic changes in the human methylome during differentiation. Genome Res 2010;20:320–31. https://doi.org/10.1101/GR.101907.109.

[45] Nguyen CT, Gonzales FA, Jones PA. Altered chromatin structure associated with methylation-induced gene silencing in cancer cells: correlation of accessibility, methylation, MeCP2 binding and acetylation. Nucleic Acids Res 2001;29:4598. https://doi.org/10.1093/NAR/29.22.4598.

[46] Pappalardo XG, Barra V. Losing DNA methylation at repetitive elements and breaking bad. Epigenetics Chromatin 2021 141 2021;14:1–21. https://doi.org/10.1186/S13072-021-00400-Z.

[47] Barbosa M, Joshi RS, Garg P, Martin-Trujillo A, Patel N, Jadhav B, et al. Identification of rare de novo epigenetic variations in congenital disorders. Nat Commun 2018 91 2018;9:1–11. https://doi.org/10.1038/s41467-018-04540-x.

[48] Kim YI. Nutritional epigenetics: impact of folate deficiency on DNA methylation and colon cancer susceptibility. J Nutr 2005;135:2703–9. https://doi.org/10.1093/JN/135.11.2703.

[49] Naeini MM, Ardekani AM. Noncoding RNAs and Cancer. Avicenna J Med Biotechnol 2009;1:55.

[50] Kulis M, Esteller M. DNA methylation and cancer. Adv Genet 2010;70:27–56. https://doi.org/10.1016/B978-0-12-380866-0.60002-2.

[51] Yoo CB, Jones PA. Epigenetic therapy of cancer: past, present and future. Nat Rev Drug Discov 2006 51 2006;5:37–50. https://doi.org/10.1038/nrd1930.

[52] Christman JK. 5-Azacytidine and 5-aza-2'-deoxycytidine as inhibitors of DNA methylation: mechanistic studies and their implications for cancer therapy. Oncogene 2002;21:5483–95. https://doi.org/10.1038/SJ.ONC.1205699.

[53] Kaminskas E, Farrell A, Abraham S, Baird A, Hsieh LS, Lee SL, et al. Approval summary: azacitidine for treatment of myelodysplastic syndrome subtypes. Clin Cancer Res 2005;11:3604–8. https://doi.org/10.1158/1078-0432.CCR-04-2135.

[54] Fece de la Cruz F, Corcoran RB. Methylation in cell-free DNA for early cancer detection. Ann Oncol 2018;29:1351–3. https://doi.org/10.1093/annonc/mdy134.

[55] Zhang Q, Hu G, Yang Q, Dong R, Xie X, Ma D, et al. A multiplex methylation-specific PCR assay for the detection of early-stage ovarian cancer using cell-free serum DNA. Gynecol Oncol 2013;130:132–9. https://doi.org/10.1016/J.YGYNO.2013.04.048.

[56] Greer JM, Mccombe PA. The role of epigenetic mechanisms and processes in autoimmune disorders. Biologics 2012;6:307. https://doi.org/10.2147/BTT.S24067.

[57] Huber LC, Brock M, Hemmatazad H, Giger OT, Moritz F, Trenkmann M, et al. Histone deacetylase/acetylase activity in total synovial tissue derived from rheumatoid arthritis and osteoarthritis patients. Arthritis Rheum 2007;56:1087–93. https://doi.org/10.1002/ART.22512.

[58] Mastronardi FG, Noor A, Wood DD, Paton T, Moscarello MA. Peptidyl argininedeiminase 2 CpG island in multiple sclerosis white matter is hypomethylated. J Neurosci Res 2007;85:2006–16. https://doi.org/10.1002/JNR.21329.

[59] Day JJ, Sweatt JD. DNA methylation and memory formation. Nat Neurosci 2010;13:1319–23. https://doi.org/10.1038/NN.2666.

[60] Coda DM, Gräff J. Neurogenetic and Neuroepigenetic Mechanisms in Cognitive Health and Disease. Front Mol Neurosci 2020;13:205. https://doi.org/10.3389/FNMOL.2020.589109/BIBTEX.

[61] Siegmund KD, Connor CM, Campan M, Long TL, Weisenberger DJ, Biniszkiewicz D, et al. DNA Methylation in the Human Cerebral Cortex Is Dynamically Regulated throughout the Life Span and Involves Differentiated Neurons. PLoS One 2007;2. https://doi.org/10.1371/JOURNAL.PONE.0000895.

[62] Zhang Y, Zeng C. Role of DNA methylation in cardiovascular diseases. Clin Exp Hypertens 2016;38:261–7. https://doi.org/10.3109/10641963.2015.1107087.

[63] Rivière G, Lienhard D, Andrieu T, Vieau D, Frey BM, Frey FJ. Epigenetic regulation of somatic angiotensin-converting enzyme by DNA methylation and histone acetylation. Epigenetics 2011;6:478–89. https://doi.org/10.4161/EPI.6.4.14961.

[64] Gonzalez-Jaramillo V, Portilla-Fernandez E, Glisic M, Voortman

T, Bramer W, Chowdhury R, et al. The role of DNA methylation and histone modifications in blood pressure: a systematic review. J Hum Hypertens 2019;33:703–15. https://doi.org/10.1038/S41371-019-0218-7.

[65] Richard MA, Huan T, Ligthart S, Gondalia R, Jhun MA, Brody JA, et al. DNA Methylation Analysis Identifies Loci for Blood Pressure Regulation. Am J Hum Genet 2017;101:888–902. https://doi.org/10.1016/J.AJHG.2017.09.028.

[66] Syme C, Shin J, Richer L, Gaudet D, Fornage M, Paus T, et al. Epigenetic Loci of Blood Pressure. Circ Genomic Precis Med 2019;12:E002341. https://doi.org/10.1161/CIRCGEN.118.002341.

[67] Samblas M, Milagro FI, Martínez A. DNA methylation markers in obesity, metabolic syndrome, and weight loss. Epigenetics 2019;14:421. https://doi.org/10.1080/15592294.2019.1595297.

[68] Aslibekyan S, Demerath EW, Mendelson M, Zhi D, Guan W, Liang L, et al. Epigenome-wide study identifies novel methylation loci associated with body mass index and waist circumference. Obesity (Silver Spring) 2015;23:1493–501. https://doi.org/10.1002/OBY.21111.

[69] Sharp GC, Lawlor DA, Richmond RC, Fraser A, Simpkin A, Suderman M, et al. Maternal pre-pregnancy BMI and gestational weight gain, offspring DNA methylation and later offspring adiposity: findings from the Avon Longitudinal Study of Parents and Children. Int J Epidemiol 2015;44:1288–304. https://doi.org/10.1093/IJE/DYV042.

[70] Melnikova I. Rare diseases and orphan drugs. Nat Rev Drug Discov 2012;11:267–8. https://doi.org/10.1038/NRD3654.

[71] Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved diagnostic yield compared with targeted gene sequencing panels suggests a role for whole-genome sequencing as a first-tier genetic test. Genet Med 2018;20:435–43. https://doi.org/10.1038/GIM.2017.119.

[72] Aref-Eshghi E, Bend EG, Colaiacovo S, Caudle M, Chakrabarti R, Napier M, et al. Diagnostic Utility of Genome-wide DNA Methylation Testing in Genetically Unsolved Individuals with Suspected Hereditary Conditions. Am J Hum Genet 2019;104:685–700. https://doi.org/10.1016/J.AJHG.2019.03.008.

[73] Garg P, Jadhav B, Rodriguez OL, Patel N, Martin-Trujillo A, Jain M, et al. A Survey of Rare Epigenetic Variation in 23,116 Human

Genomes Identifies Disease-Relevant Epivariations and CGG Expansions. Am J Hum Genet 2020;107:654–69. https://doi.org/10.1016/J.AJHG.2020.08.019.

[74] Campagna MP, Xavier A, Lechner-Scott J, Maltby V, Scott RJ, Butzkueven H, et al. Epigenome-wide association studies: current knowledge, strategies and recommendations. Clin Epigenetics 2021;13:1–24. https://doi.org/10.1186/S13148-021-01200-8/FIGURES/12.

[75] Laird PW. Principles and challenges of genomewide DNA methylation analysis. Nat Rev Genet 2010;11:191–203. https://doi.org/10.1038/NRG2732.

[76] Tompa R, McCallum CM, Delrow J, Henikoff JG, Van Steensel B, Henikoff S. Genome-wide profiling of DNA methylation reveals transposon targets of CHROMOMETHYLASE3. Curr Biol 2002;12:65–8. https://doi.org/10.1016/S0960-9822(01)00622-4.

[77] Singer BD. A Practical Guide to the Measurement and Analysis of DNA Methylation. Am J Respir Cell Mol Biol 2019;61:417–28. https://doi.org/10.1165/RCMB.2019-0150TR.

[78] Hayatsu H. Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis--a personal account. Proc Jpn Acad Ser B Phys Biol Sci 2008;84:321–30. https://doi.org/10.2183/PJAB.84.321.

[79] New England Biolabs. Key restriction enzymes for DNA methylation analysis n.d. https://doi.org/10.1186/gb-2010-12-6-116.

[80] Lee JR, Ryu DS, Park SJ, Choe SH, Cho HM, Lee SR, et al. Successful application of human-based methyl capture sequencing for methylome analysis in non-human primate models. BMC Genomics 2018;19. https://doi.org/10.1186/S12864-018-4666-1.

[81] Epigentek. DNA Bisulfite Conversion n.d. https://www.epigentek.com/.

[82] Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox A V., et al. DNA Methylation Profiling of the Human Major Histocompatibility Complex: A Pilot Study for the Human Epigenome Project. PLoS Biol 2004;2. https://doi.org/10.1371/JOURNAL.PBIO.0020405.

[83] Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, et al. DNA methylation patterns associate with genetic and

gene expression variation in HapMap cell lines. Genome Biol 2011;12:1–13. https://doi.org/10.1186/GB-2011-12-1-R10/FIGURES/5.

[84] Zhang D, Cheng L, Badner JA, Chen C, Chen Q, Luo W, et al. Genetic control of individual differences in gene-specific methylation in human brain. Am J Hum Genet 2010;86:411–9. https://doi.org/10.1016/J.AJHG.2010.02.005.

[85] Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet 2021 539 2021;53:1311–21. https://doi.org/10.1038/s41588-021-00923-x.

[86] Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun 2014;5:3365. https://doi.org/10.1038/NCOMMS4365.

[87] Milunsky A, Milunsky JM. Genetic Disorders and the Fetus: Diagnosis, Prevention, and Treatment. Wiley; 2015.

[88] Anton E, Blanco J, Egozcue J, Vidal F. Sperm FISH studies in seven male carriers of Robertsonian translocation t(13;14)(q10;q10). Hum Reprod 2004;19:1345–51. https://doi.org/10.1093/HUMREP/DEH232.

[89] McCartney DL, Walker RM, Morris SW, Anderson SM, Duff BJ, Marioni RE, et al. Altered DNA methylation associated with a translocation linked to major mental illness. Npj Schizophr 2018 41 2018;4:1–7. https://doi.org/10.1038/s41537-018-0047-7.

[90] Cáceres A, González JR. Following the footprints of polymorphic inversions on SNP data: From detection to association tests. Nucleic Acids Res 2015;43:[e53]. https://doi.org/10.1093/nar/gkv073.

[91] González JR, Ruiz-Arenas C, Cáceres A, Morán I, López-Sánchez M, Alonso L, et al. Polymorphic Inversions Underlie the Shared Genetic Susceptibility of Obesity-Related Diseases. Am J Hum Genet 2020;106:846–58. https://doi.org/10.1016/j.ajhg.2020.04.017.

[92] Ruiz-Arenas C, Cáceres A, Moreno V, González JR. Common polymorphic inversions at 17q21.31 and 8p23.1 associate with cancer prognosis. Hum Genomics 2019;13:[57]. https://doi.org/10.1186/s40246-019-0242-2.

[93] Tantisira KG, Lazarus R, Litonjua AA, Klanderman B, Weiss ST. Chromosome 17: Association of a large inversion polymorphism with corticosteroid response in asthma. Pharmacogenet Genomics 2008;18:733–7. https://doi.org/10.1097/FPC.0b013e3282fe6ebf.

[94] González JR, Cáceres A, Esko T, Cuscó I, Puig M, Esnaola M, et al. A common 16p11.2 inversion underlies the joint susceptibility to asthma and obesity. Am J Hum Genet 2014;94:361–72. https://doi.org/10.1016/j.ajhg.2014.01.015.

[95] Luciano M, Hagenaars SP, Davies G, Hill WD, Clarke TK, Shirali M, et al. Association analysis in over 329,000 individuals identifies 116 independent variants influencing neuroticism. Nat Genet 2018;50:6–11. https://doi.org/10.1038/s41588-017-0013-8.

[96] Puig M, Casillas S, Villatoro S, Cáceres M. Human inversions and their functional consequences. Brief Funct Genomics 2015;14:369–79. https://doi.org/10.1093/bfgp/elv020.

[97] Ruiz-Arenas C, Cáceres A, López-Sánchez M, Tolosana I, Pérez-Jurado L, González JR. scoreInvHap: Inversion genotyping for genome-wide association studies. PLoS Genet 2019;15:[e1008203]. https://doi.org/10.1371/journal.pgen.1008203.

[98] Jamil MA, Sharma A, Nuesgen N, Pezeshkpoor B, Heimbach A, Pavlova A, et al. F8 inversions at Xq28 causing hemophilia a are associated with specific methylation changes: Implication for molecular epigenetic diagnosis. Front Genet 2019;10:508. https://doi.org/10.3389/fgene.2019.00508.

[99] Smith AC, Suzuki M, Thompson R, Choufani S, Higgins MJ, Chiu IW, et al. Maternal gametic transmission of translocations or inversions of human chromosome 11p15.5 results in regional DNA hypermethylation and downregulation of CDKN1C expression. Genomics 2012;99:25–35. https://doi.org/10.1016/j.ygeno.2011.10.007.

[100] MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 2014;42. https://doi.org/10.1093/NAR/GKT958.

[101] Zarrei M, MacDonald JR, Merico D, Scherer SW. A copy number variation map of the human genome. Nat Rev Genet 2015 163 2015;16:172–83. https://doi.org/10.1038/nrg3871.

[102] Shi X, Radhakrishnan S, Wen J, Chen JY, Chen J, Lam BA, et al.

Association of CNVs with methylation variation. Npj Genomic Med 2020;5:[41]. https://doi.org/10.1038/s41525-020-00145-w.

[103] Sun W, Bunn P, Jin C, Little P, Zhabotynsky V, Perou CM, et al. The association between copy number aberration, DNA methylation and gene expression in tumor samples. Nucleic Acids Res 2018;46:3009–18. https://doi.org/10.1093/NAR/GKY131.

[104] Li J, Harris RA, Cheung SW, Coarfa C, Jeong M, Goodell MA, et al. Genomic Hypomethylation in the Human Germline Associates with Selective Structural Mutability in the Human Genome. PLOS Genet 2012;8:e1002692. https://doi.org/10.1371/JOURNAL.PGEN.1002692.

[105] Hassold T, Hunt P. To err (meiotically) is human: the genesis of human aneuploidy. Nat Rev Genet 2001 24 2001;2:280–91. https://doi.org/10.1038/35066065.

[106] Muskens IS, Li S, Jackson T, Elliot N, Hansen HM, Myint SS, et al. The genome-wide impact of trisomy 21 on DNA methylation and its implications for hematopoiesis. Nat Commun 2021 121 2021;12:1–15. https://doi.org/10.1038/s41467-021-21064-z.

[107] Wild CP. Complementing the Genome with an "'Exposome'": The Outstanding Challenge of Environmental Exposure Measurement in Molecular Epidemiology 2005;14:1847–50. https://doi.org/10.1158/1055-9965.EPI-05-0456.

[108] Ashby J, Tennant RW. Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP. Mutat Res 1991;257:229–306. https://doi.org/10.1016/0165-1110(91)90003-E.

[109] Weng YI, Hsu PY, Liyanarachchi S, Liu J, Deatherage DE, Huang YW, et al. Epigenetic influences of low-dose bisphenol A in primary human breast epithelial cells. Toxicol Appl Pharmacol 2010;248:111–21. https://doi.org/10.1016/J.TAAP.2010.07.014.

[110] Prins GS, Tang WY, Belmonte J, Ho SM. Perinatal exposure to oestradiol and bisphenol A alters the prostate epigenome and increases susceptibility to carcinogenesis. Basic Clin Pharmacol Toxicol 2008;102:134–8. https://doi.org/10.1111/J.1742-7843.2007.00166.X.

[111] Ellen TP, Kluz T, Harder ME, Xiong J, Costa M. Heterochromatinization as a Potential Mechanism of Nickel-Induced Carcinogenesis. Biochemistry 2009;48:4626. https://doi.org/10.1021/BI900246H.

[112] Takiguchi M, Achanzar WE, Qu W, Li G, Waalkes MP. Effects of cadmium on DNA-(Cytosine-5) methyltransferase activity and DNA methylation status during cadmium-induced cellular transformation. Exp Cell Res 2003;286:355–65. https://doi.org/10.1016/S0014-4827(03)00062-4.

[113] Ren X, Mchale CM, Skibola CF, Smith AH, Smith MT, Zhang L. An emerging role for epigenetic dysregulation in arsenic toxicity and carcinogenesis. Environ Health Perspect 2011;119:11–9. https://doi.org/10.1289/EHP.1002114.

[114] Barrès R, Yan J, Egan B, Treebak JT, Rasmussen M, Fritz T, et al. Acute exercise remodels promoter methylation in human skeletal muscle. Cell Metab 2012;15:405–11. https://doi.org/10.1016/J.CMET.2012.01.001.

[115] Nakajima K, Takeoka M, Mori M, Hashimoto S, Sakurai A, Nose H, et al. Exercise effects on methylation of ASC gene. Int J Sports Med 2010;31:671–5. https://doi.org/10.1055/S-0029-1246140.

[116] Zeng H, Irwin ML, Lu L, Risch H, Mayne S, Mu L, et al. Physical activity and breast cancer survival: an epigenetic link through reduced methylation of a tumor suppressor gene L3MBTL1. Breast Cancer Res Treat 2012;133:127–35. https://doi.org/10.1007/S10549-011-1716-7.

[117] Ntanasis-Stathopoulos J, Tzanninis JG, Philippou A, Koutsilieris M. Epigenetic regulation on gene expression induced by physical exercise. J Musculoskelet Neuronal Interact 2013.

[118] Mehta D, Klengel T, Conneely KN, Smith AK, Altmann A, Pace TW, et al. Childhood maltreatment is associated with distinct genomic and epigenetic profiles in posttraumatic stress disorder. Proc Natl Acad Sci U S A 2013;110:8302–7. https://doi.org/10.1073/PNAS.1217750110.

[119] Khashan AS, Abel KM, McNamee R, Pedersen MG, Webb RT, Baker PN, et al. Higher risk of offspring schizophrenia following antenatal maternal exposure to severe adverse life events. Arch Gen Psychiatry 2008;65:146–52. https://doi.org/10.1001/ARCHGENPSYCHIATRY.2007.20.

[120] Lahtinen A, Puttonen S, Vanttola P, Viitasalo K, Sulkava S, Pervjakova N, et al. A distinctive DNA methylation pattern in insufficient sleep. Sci Reports 2019 91 2019;9:1–9. https://doi.org/10.1038/s41598-018-38009-0.

[121] Dugué PA, Jung CH, Joo JE, Wang X, Wong EM, Makalic E, et

al. Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility. Epigenetics 2020;15:358–68. https://doi.org/10.1080/15592294.2019.1668739.

[122] Zeilinger S, Kühnel B, Klopp N, Baurecht H, Kleinschmidt A, Gieger C, et al. Tobacco smoking leads to extensive genome-wide changes in DNA methylation. PLoS One 2013;8. https://doi.org/10.1371/JOURNAL.PONE.0063812.

[123] Dogan M V., Beach SRH, Philibert RA. Genetically contextual effects of smoking on genome wide DNA methylation. Am J Med Genet B Neuropsychiatr Genet 2017;174:595–607. https://doi.org/10.1002/AJMG.B.32565.

[124] Christiansen C, Castillo-Fernandez JE, Domingo-Relloso A, Zhao W, El-Sayed Moustafa JS, Tsai PC, et al. Novel DNA methylation signatures of tobacco smoking with trans-ethnic effects. Clin Epigenetics 2021;13. https://doi.org/10.1186/S13148-021-01018-4.

[125] Sikdar S, Joehanes R, Joubert BR, Xu CJ, Vives-Usano M, Rezwan FI, et al. Comparison of smoking-related DNA methylation between newborns from prenatal exposure and adults from personal smoking. Epigenomics 2019;11:1487–500. https://doi.org/10.2217/EPI-2019-0066.

[126] Xu K, Montalvo-Ortiz JL, Zhang X, Southwick SM, Krystal JH, Pietrzak RH, et al. Epigenome-Wide DNA Methylation Association Analysis Identified Novel Loci in Peripheral Cells for Alcohol Consumption Among European American Male Veterans. Alcohol Clin Exp Res 2019;43:2111–21. https://doi.org/10.1111/ACER.14168.

[127] Toinét Cronjé H, Elliott HR, Nienaber-Rousseau C, Pieters M. Replication and expansion of epigenome-wide association literature in a black South African population. Clin Epigenetics 2020;12. https://doi.org/10.1186/S13148-019-0805-Z.

[128] Lohoff FW, Clarke TK, Kaminsky ZA, Walker RM, Bermingham ML, Jung J, et al. Epigenome-wide association study of alcohol consumption in N = 8161 individuals and relevance to alcohol use disorder pathophysiology: identification of the cystine/glutamate transporter SLC7A11 as a top target. Mol Psychiatry 2022;27. https://doi.org/10.1038/S41380-021-01378-6.

[129] Dugué PA, Wilson R, Lehne B, Jayasekara H, Wang X, Jung CH, et al. Alcohol consumption is associated with widespread changes in blood DNA methylation: Analysis of cross-sectional and

longitudinal data. Addict Biol 2021;26. https://doi.org/10.1111/ADB.12855.

[130] Stephenson M, Bollepalli S, Cazaly E, Salvatore JE, Barr P, Rose RJ, et al. Associations of Alcohol Consumption With Epigenome-Wide DNA Methylation and Epigenetic Age Acceleration: Individual-Level and Co-twin Comparison Analyses. Alcohol Clin Exp Res 2021;45:318–28. https://doi.org/10.1111/ACER.14528.

[131] Osborne AJ, Pearson JF, Noble AJ, Gemmell NJ, Horwood LJ, Boden JM, et al. Genome-wide DNA methylation analysis of heavy cannabis exposure in a New Zealand longitudinal cohort. Transl Psychiatry 2020 101 2020;10:1–10. https://doi.org/10.1038/s41398-020-0800-3.

[132] Markunas CA, Hancock DB, Xu Z, Quach BC, Fang F, Sandler DP, et al. Epigenome-wide analysis uncovers a blood-based DNA methylation biomarker of lifetime cannabis use. Am J Med Genet B Neuropsychiatr Genet 2021;186:173–82. https://doi.org/10.1002/AJMG.B.32813.

[133] Heijmans BT, Tobi EW, Stein AD, Putter H, Blauw GJ, Susser ES, et al. Persistent epigenetic differences associated with prenatal exposure to famine in humans. Proc Natl Acad Sci U S A 2008;105:17046–9. https://doi.org/10.1073/PNAS.0806560105.

[134] Stidley CA, Picchi MA, Leng S, Willink R, Crowell RE, Flores KG, et al. Multivitamins, folate, and green vegetables protect against gene promoter methylation in the aerodigestive tract of smokers. Cancer Res 2010;70:568–74. https://doi.org/10.1158/0008-5472.CAN-09-3410.

[135] Ono H, Iwasaki M, Kuchiba A, Kasuga Y, Yokoyama S, Onuma H, et al. Association of dietary and genetic factors related to one-carbon metabolism with global methylation level of leukocyte DNA. Cancer Sci 2012;103:2159–64. https://doi.org/10.1111/CAS.12013.

[136] Zhang FF, Santella RM, Wolff M, Kappil MA, Markowitz SB, Morabia A. White blood cell global methylation and IL-6 promoter methylation in association with diet and lifestyle risk factors in a cancer-free population. Epigenetics 2012;7:606–14. https://doi.org/10.4161/EPI.20236.

[137] Agodi A, Barchitta M, Quattrocchi A, Maugeri A, Canto C, Marchese AE, et al. Low fruit consumption and folate deficiency are associated with LINE-1 hypomethylation in women of a

cancer-free population. Genes Nutr 2015;10. https://doi.org/10.1007/S12263-015-0480-4.

[138] Barchitta M, Maugeri A, Quattrocchi A, Barone G, Mazzoleni P, Catalfo A, et al. Mediterranean Diet and Particulate Matter Exposure Are Associated With LINE-1 Methylation: Results From a Cross-Sectional Study in Women. Front Genet 2018;9. https://doi.org/10.3389/FGENE.2018.00514.

[139] Nicodemus-Johnson J, Sinnott RA. Fruit and Juice Epigenetic Signatures Are Associated with Independent Immunoregulatory Pathways. Nutrients 2017;9. https://doi.org/10.3390/NU9070752.

[140] Mukherjee S, Dasgupta S, Mishra PK, Chaudhury K. Air pollution-induced epigenetic changes: disease development and a possible link with hypersensitivity pneumonitis. Environ Sci Pollut Res Int 2021;28:55981. https://doi.org/10.1007/S11356-021-16056-X.

[141] De Nys S, Duca RC, Nawrot T, Hoet P, Van Meerbeek B, Van Landuyt KL, et al. Temporal variability of global DNA methylation and hydroxymethylation in buccal cells of healthy adults: Association with air pollution. Environ Int 2018;111:301–8. https://doi.org/10.1016/J.ENVINT.2017.11.002.

[142] Janssen BG, Godderis L, Pieters N, Poels K, Kiciński M, Cuypers A, et al. Placental DNA hypomethylation in association with particulate air pollution in early life. Part Fibre Toxicol 2013;10. https://doi.org/10.1186/1743-8977-10-22.

[143] Soberanes S, Gonzalez A, Urich D, Chiarella SE, Radigan KA, Osornio-Vargas A, et al. Particulate matter Air Pollution induces hypermethylation of the p16 promoter Via a mitochondrial ROS-JNK-DNMT1 pathway. Sci Reports 2012 21 2012;2:1–8. https://doi.org/10.1038/srep00275.

[144] Breton C V., Salam MT, Wang X, Byun HM, Siegmund KD, Gilliland FD. Particulate matter, DNA methylation in nitric oxide synthase, and childhood respiratory disease. Environ Health Perspect 2012;120:1320–6. https://doi.org/10.1289/EHP.1104439.

[145] Prunicki M, Stell L, Dinakarpandian D, de Planell-Saguer M, Lucas RW, Hammond SK, et al. Exposure to NO 2, CO, and PM 2.5 is linked to regional DNA methylation differences in asthma. Clin Epigenetics 2018;10:2. https://doi.org/10.1186/S13148-017-0433-4.

[146] Padmanabhan JL, Shah JL, Tandon N, Keshavan MS. The

"polyenviromic risk score": Aggregating environmental risk factors predicts conversion to psychosis in familial high-risk subjects. Schizophr Res 2017;181:17–22. https://doi.org/10.1016/J.SCHRES.2016.10.014.

[147] Jeon EJ, Kang SH, Piao YH, Kim SW, Kim JJ, Lee BJ, et al. Development of the Korea-Polyenvironmental Risk Score for Psychosis. Psychiatry Investig 2022;19:197–206. https://doi.org/10.30773/PI.2021.0328.

[148] Oliver D, Radua J, Reichenberg A, Uher R, Fusar-Poli P. Psychosis Polyrisk Score (PPS) for the Detection of Individuals At-Risk and the Prediction of Their Outcomes. Front Psychiatry 2019;10. https://doi.org/10.3389/FPSYT.2019.00174.

[149] Vassos E, Sham P, Kempton M, Trotta A, Stilo SA, Gayer-Anderson C, et al. The Maudsley environmental risk score for psychosis. Psychol Med 2020;50:2213–20. https://doi.org/10.1017/S0033291719002319.

[150] Ellulu MS, Jalambo MO. Gene-environment Interaction: The Causes of High Obesity Incidence. Kathmandu Univ Med J (KUMJ) 2017;15:91–3. https://doi.org/10.15761/mca.1000104.

[151] Caspi A, McCray J, Moffitt TE, Mill J, Martin J, Craig IW, et al. Role of genotype in the cycle of violence in maltreated children. Science 2002;297:851–4. https://doi.org/10.1126/SCIENCE.1072290.

[152] Caspi A, Sugden K, Moffitt TE, Taylor A, Craig IW, Harrington HL, et al. Influence of life stress on depression: moderation by a polymorphism in the 5-HTT gene. Science 2003;301:386–9. https://doi.org/10.1126/SCIENCE.1083968.

[153] Manuck SB, McCaffery JM. Gene-environment interaction. Annu Rev Psychol 2014;65:41–70. https://doi.org/10.1146/ANNUREV-PSYCH-010213-115100.

[154] Uher R. The role of genetic variation in the causation of mental illness: an evolution-informed framework. Mol Psychiatry 2009;14:1072–82. https://doi.org/10.1038/MP.2009.85.

[155] Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet 2003 333 2003;33:245–54. https://doi.org/10.1038/ng1089.

[156] Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics 2020;15:1.

https://doi.org/10.1080/15592294.2019.1644879.

[157] Law PP, Holland ML. DNA methylation at the crossroads of gene and environment interactions. Essays Biochem 2019;63:717. https://doi.org/10.1042/EBC20190031.

[158] Czamara D, Eraslan G, Page CM, Lahti J, Lahti-Pulkkinen M, Hämäläinen E, et al. Integrated analysis of environmental and genetic influences on cord blood DNA methylation in new-borns. Nat Commun 2019 101 2019;10:1–18. https://doi.org/10.1038/s41467-019-10461-0.

[159] Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): Project rationale and design. Environ Health Perspect 2014;122:535–44. https://doi.org/10.1289/ehp.1307204.

[160] Wright J, Small N, Raynor P, Tuffnell D, Bhopal R, Cameron N, et al. Cohort profile: The born in bradford multi-ethnic family cohort study. Int J Epidemiol 2013;42:978–91. https://doi.org/10.1093/ije/dys112.

[161] Heude B, Forhan A, Slama R, Douhaud L, Bedel S, Saurel-Cubizolles MJ, 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. https://doi.org/10.1093/ije/dyv151.

[162] Guxens M, Ballester F, Espada M, Fernández 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. https://doi.org/10.1093/ije/dyr054.

[163] Grazuleviciene R, Danileviciute A, Dedele A, Vencloviene J, Andrusaityte S, Uždanaviciute I, et al. Surrounding greenness, proximity to city parks and pregnancy outcomes in Kaunas cohort study. Int J Hyg Environ Health 2015;218:358–65. https://doi.org/10.1016/j.ijheh.2015.02.004.

[164] Magnus P, Birke C, Vejrup K, Haugan A, Alsaker E, Daltveit AK, et al. Cohort Profile Update: The Norwegian Mother and Child Cohort Study (MoBa). Int J Epidemiol 2016;45:382–8. https://doi.org/10.1093/ije/dyw029.

[165] 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;46:1392–3. https://doi.org/10.1093/ije/dyx084.

[166] Tamayo-Uria I, Maitre L, Thomsen C, Nieuwenhuijsen MJ, Chatzi L, Siroux V, et al. The early-life exposome: Description and patterns in six European countries. Environ Int 2019;123:189–200. https://doi.org/10.1016/j.envint.2018.11.067.

[167] TruDiagnostic. TruAge Complete Collection – TruDiagnostic™ n.d. https://trudiagnostic.com/products/truage-complete-epigenetic-collection (accessed October 17, 2022).

[168] Ruiz-Arenas C. A multi-omics approach improves diagnosis in major isolated congenital heart disease. ASHG Virtual Meet 2020.

[169] Enokida H, Shiina H, Urakami S, Terashima M, Ogishima T, Li LC, et al. Smoking influences aberrant CpG hypermethylation of multiple genes in human prostate carcinoma. Cancer 2006;106:79–86. https://doi.org/10.1002/CNCR.21577.

[170] Monick MM, Beach SRH, Plume J, Sears R, Gerrard M, Brody GH, et al. Coordinated Changes in AHRR Methylation in Lymphoblasts and Pulmonary Macrophages from Smokers. Am J Med Genet 2012;159B:141. https://doi.org/10.1002/AJMG.B.32021.

[171] Wan ES, Qiu W, Baccarelli A, Carey VJ, Bacherman H, Rennard SI, et al. Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome. Hum Mol Genet 2012;21:3073–82. https://doi.org/10.1093/HMG/DDS135.

[172] Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environ Health Perspect 2012;120:1425–31. https://doi.org/10.1289/EHP.1205412.

[173] Shenker NS, Polidoro S, van Veldhoven K, Sacerdote C, Ricceri F, Birrell MA, et al. Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking. Hum Mol Genet 2013;22:843–51. https://doi.org/10.1093/HMG/DDS488.

[174] Battram T, Yousefi P, Crawford G, Prince C, Sheikhali Babaei M, Sharp G, et al. The EWAS Catalog: a database of epigenome-wide association studies. Wellcome Open Res 2022;7:41. https://doi.org/10.12688/WELLCOMEOPENRES.17598.2.

[175] Liang X, Justice AC, So-Armah K, Krystal JH, Sinha R, Xu K.

DNA methylation signature on phosphatidylethanol, not on self-reported alcohol consumption, predicts hazardous alcohol consumption in two distinct populations. Mol Psychiatry 2021;26:2238–53. https://doi.org/10.1038/S41380-020-0668-X.

[176] Wilson LE, Xu Z, Harlid S, White AJ, Troester MA, Sandler DP, et al. Alcohol and DNA Methylation: An Epigenome-Wide Association Study in Blood and Normal Breast Tissue. Am J Epidemiol 2019;188:1055–65. https://doi.org/10.1093/AJE/KWZ032.

[177] Liu C, Marioni RE, Hedman AK, Pfeiffer L, Tsai PC, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Mol Psychiatry 2018 232 2016;23:422–33. https://doi.org/10.1038/mp.2016.192.

[178] Watson CT, Szutorisz H, Garg P, Martin Q, Landry JA, Sharp AJ, et al. Genome-Wide DNA Methylation Profiling Reveals Epigenetic Changes in the Rat Nucleus Accumbens Associated With Cross-Generational Effects of Adolescent THC Exposure. Neuropsychopharmacology 2015;40:2993–3005. https://doi.org/10.1038/NPP.2015.155.

[179] Aragona J, Lee CK. Scoliosis in fetal alcohol syndrome: A case report. Orthopedics 1981;4:1141–3. https://doi.org/10.3928/0147-7447-19811001-06.

[180] Gallagher C, McCarthy FP, Ryan RM, Khashan AS. Maternal Alcohol Consumption During Pregnancy and the Risk of Autism Spectrum Disorders in Offspring: A Retrospective Analysis of the Millennium Cohort Study. J Autism Dev Disord 2018;48:3773. https://doi.org/10.1007/S10803-018-3626-6.

[181] Singer AB, Aylsworth AS, Cordero C, Croen LA, DiGuiseppi C, Fallin MD, et al. Prenatal Alcohol Exposure in Relation to Autism Spectrum Disorder: Findings from the Study to Explore Early Development (SEED). Paediatr Perinat Epidemiol 2017;31:573. https://doi.org/10.1111/PPE.12404.

[182] Virdis A, Giannarelli C, Fritsch Neves M, Taddei S, Ghiadoni L. Cigarette smoking and hypertension. Curr Pharm Des 2010;16:2518–25. https://doi.org/10.2174/138161210792062920.

[183] Dikalov S, Itani H, Richmond B, Vergeade A, Jamshedur Rahman SM, Boutaud O, et al. Tobacco smoking induces cardiovascular mitochondrial oxidative stress, promotes endothelial dysfunction, and enhances hypertension. Am J Physiol Heart Circ Physiol

2019;316:H639–46.
https://doi.org/10.1152/AJPHEART.00595.2018.

[184] Minzer S, Losno RA, Casas R. The Effect of Alcohol on Cardiovascular Risk Factors: Is There New Information? Nutrients 2020;12:1–22. https://doi.org/10.3390/NU12040912.

[185] Tasnim S, Tang C, Musini VM, Wright JM. Effect of alcohol on blood pressure. Cochrane Database Syst Rev 2020;2020. https://doi.org/10.1002/14651858.CD012787.PUB2.

[186] Abuhasira R, Haviv YS, Leiba M, Leiba A, Ryvo L, Novack V. Cannabis is associated with blood pressure reduction in older adults - A 24-hours ambulatory blood pressure monitoring study. Eur J Intern Med 2021;86:79–85. https://doi.org/10.1016/J.EJIM.2021.01.005.

[187] Alshaarawy O, Elbaz HA. Cannabis use and blood pressure levels: United States National Health and Nutrition Examination Survey, 2005-2012. J Hypertens 2016;34:1507–12. https://doi.org/10.1097/HJH.0000000000000990.

[188] Jones RT. Cardiovascular system effects of marijuana. J Clin Pharmacol 2002;42. https://doi.org/10.1002/J.1552-4604.2002.TB06004.X.

[189] O'Keefe JH, Bhatti SK, Bajwa A, DiNicolantonio JJ, Lavie CJ. Alcohol and cardiovascular health: the dose makes the poison…or the remedy. Mayo Clin Proc 2014;89:382–93. https://doi.org/10.1016/J.MAYOCP.2013.11.005.

[190] Giner-Delgado C, Villatoro S, Lerga-Jaso J, Gayà-Vidal M, Oliva M, Castellano D, et al. Evolutionary and functional impact of common polymorphic inversions in the human genome. Nat Commun 2019;10:[4222]. https://doi.org/10.1038/s41467-019-12173-x.

[191] Salm MPA, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res 2012;22:1144–53. https://doi.org/10.1101/gr.126037.111.

[192] de Jong S, Chepelev I, Janson E, Strengman E, van den Berg LH, Veldink JH, et al. Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics 2012;13:458. https://doi.org/10.1186/1471-2164-13-458.

[193] Shanta O, Noor A, Chaisson MJP, Sanders AD, Zhao X, Malhotra A, et al. The effects of common structural variants on 3D chromatin structure. BMC Genomics 2020;21:1–10. https://doi.org/10.1186/s12864-020-6516-1.

[194] Schlauch KA, Read RW, Lombardi VC, Elhanan G, Metcalf WJ, Slonim AD, et al. A Comprehensive Genome-Wide and Phenome-Wide Examination of BMI and Obesity in a Northern Nevadan Cohort. G3 2020;10:645–64. https://doi.org/10.1534/G3.119.400910.

[195] MH R, A S-A, PJ S, L A. Is there a relationship between red or processed meat intake and obesity? A systematic review and meta-analysis of observational studies. Obes Rev 2014;15:740–8. https://doi.org/10.1111/OBR.12172.

[196] You W, Henneberg M. Meat consumption providing a surplus energy in modern diet contributes to obesity prevalence: an ecological analysis. BMC Nutr 2016 21 2016;2:1–11. https://doi.org/10.1186/S40795-016-0063-9.

[197] Serra-Juhé C, Cuscó I, Homs A, Flores R, Torán N, Pérez-Jurado LA. DNA methylation abnormalities in congenital heart disease. Epigenetics 2015;10:167–77. https://doi.org/10.1080/15592294.2014.998536.

[198] Yang Y-Q, Wang J, Liu X-Y, Chen X-Z, Zhang W, Wang X-Z. Mutation spectrum of GATA4 associated with congenital atrial septal defects. Arch Med Sci 2013;9:976. https://doi.org/10.5114/AOMS.2013.39788.

[199] Jiang Y, Zheng W. Cardiovascular Toxicities Upon Manganese Exposure. Cardiovasc Toxicol 2005;5:345.

[200] M S, A H-T, C G, RM H, TH S. Smoking and asthma. J Am Board Fam Med 2011;24:313–22. https://doi.org/10.3122/JABFM.2011.03.100180.

[201] A Z. Maternal smoking in pregnancy and its influence on childhood asthma. ERJ Open Res 2016;2:00042–2016. https://doi.org/10.1183/23120541.00042-2016.

[202] L Q, L Y, J M, Z M, L C, L Y, et al. The impact of PM2.5 on lung function in adults with asthma. Int J Tuberc Lung Dis 2020;24:570–6. https://doi.org/10.5588/IJTLD.19.0394.

[203] JA C, Y H, J L, KR M, NL T, K Z, et al. Methylome Analysis in Chickens Immunized with Infectious Laryngotracheitis Vaccine.

PLoS One 2015;10:[e0100476]. https://doi.org/10.1371/JOURNAL.PONE.0100476.

[204] J L, WG B, BW K. Genome-wide host responses against infectious laryngotracheitis virus vaccine infection in chicken embryo lung cells. BMC Genomics 2012;13:[143]. https://doi.org/10.1186/1471-2164-13-143.

[205] Voerman E, Gaillard R, Geurtsen ML, Jaddoe VWV. Maternal First-Trimester Cow-Milk Intake Is Positively Associated with Childhood General and Abdominal Visceral Fat Mass and Lean Mass but Not with Other Cardiometabolic Risk Factors at the Age of 10 Years. J Nutr 2021;151:1965–75. https://doi.org/10.1093/JN/NXAB047.

[206] Hrolfsdottir L, Rytter D, Hammer Bech B, Brink Henriksen T, Danielsen I, Steingrimsdottir L, et al. Maternal milk consumption, birth size and adult height of offspring: a prospective cohort study with 20 years of follow-up. Eur J Clin Nutr 2013 6710 2013;67:1036–41. https://doi.org/10.1038/ejcn.2013.151.

[207] Leary SD, Ness A, Emmett P, Smith GD. Maternal diet in pregnancy and offspring height, sitting height, and leg length. J Epidemiol Community Health 2005;59:467–72. https://doi.org/10.1136/JECH.2004.029884.

[208] Oken E, Levitan EB, Gillman MW. Maternal smoking during pregnancy and child overweight: systematic review and meta-analysis. Int J Obes (Lond) 2008;32:201–10. https://doi.org/10.1038/SJ.IJO.0803760.

[209] Møller SE, Ajslev TA, Andersen CS, Dalgård C, Sørensen TIA. Risk of childhood overweight after exposure to tobacco smoking in prenatal and early postnatal life. PLoS One 2014;9. https://doi.org/10.1371/JOURNAL.PONE.0109184.

[210] Hirai H, Okamoto S, Masuzaki H, Murata T, Ogata Y, Sato A, et al. Maternal Urinary Cotinine Concentrations During Pregnancy Predict Infant BMI Trajectory After Birth: Analysis of 89617 Mother-Infant Pairs in the Japan Environment and Children's Study. Front Endocrinol (Lausanne) 2022;13. https://doi.org/10.3389/FENDO.2022.850784.

[211] Suzuki K, Kondo N, Sato M, Tanaka T, Ando D, Yamagata Z. Gender differences in the association between maternal smoking during pregnancy and childhood growth trajectories: multilevel analysis. Int J Obes (Lond) 2011;35:53–9.

https://doi.org/10.1038/IJO.2010.198.

[212] Jia P, Cao X, Yang H, Dai S, He P, Huang G, et al. Green space access in the neighbourhood and childhood obesity. Obes Rev 2021;22 Suppl 1. https://doi.org/10.1111/OBR.13100.

[213] Lachowycz K, Jones AP. Greenspace and obesity: a systematic review of the evidence. Obes Rev 2011;12. https://doi.org/10.1111/J.1467-789X.2010.00827.X.

[214] de Bont J, Márquez S, Fernández-Barrés S, Warembourg C, Koch S, Persavento C, et al. Urban environment and obesity and weight-related behaviours in primary school children. Environ Int 2021;155. https://doi.org/10.1016/J.ENVINT.2021.106700.

[215] Okobi OE, Ajayi OO, Okobi TJ, Anaya IC, Fasehun OO, Diala CS, et al. The Burden of Obesity in the Rural Adult Population of America. Cureus 2021;13. https://doi.org/10.7759/CUREUS.15770.

[216] Miller AL, Lee HJ, Lumeng JC. Obesity-associated biomarkers and executive function in children. Pediatr Res 2015;77:143–7. https://doi.org/10.1038/PR.2014.158.

[217] Contu L, Hawkes CA. A Review of the Impact of Maternal Obesity on the Cognitive Function and Mental Health of the Offspring. Int J Mol Sci 2017;18. https://doi.org/10.3390/IJMS18051093.

[218] Edlow AG. Maternal obesity and neurodevelopmental and psychiatric disorders in offspring. Prenat Diagn 2017;37:95–110. https://doi.org/10.1002/PD.4932.

[219] Li N, Yolton K, Lanphear BP, Chen A, Kalkwarf HJ, Braun JM. Impact of Early-Life Weight Status on Cognitive Abilities in Children. Obesity (Silver Spring) 2018;26:1088–95. https://doi.org/10.1002/OBY.22192.

[220] Martin A, Booth JN, McGeown S, Niven A, Sproule J, Saunders DH, et al. Longitudinal Associations Between Childhood Obesity and Academic Achievement: Systematic Review with Focus Group Data. Curr Obes Rep 2017;6:297–313. https://doi.org/10.1007/S13679-017-0272-9.

9 | APPENDICES

9.1. PhD Portfolio

Other merits

- PIF-Salut PERIS fellowship awarded by the Departament de Salut from the Generalitat de Catalunya to develop the PhD thesis
- 1st prize in the 3-minute thesis competition during the LMB-IBM Graduate Life Science Symposium 2021
- 2nd prize in Rin4' UPF 2022, a competition to explain your PhD research in 4 minutes

  Video: http://hdl.handle.net/10230/54136
- Best poster award (1st place) in the Women in Data Science Barcelona Biomedicine 2022 Conference
- 1st prize in the Chalk Talk Competition during the 8th ISGlobal PhD Symposium

Other tasks developed during the PhD

- Pre-processing of DNA methylation data from TruDiagnostic cohort (Appendix 2)
- Processing of HELIX urine and blood samples at the laboratory AirLab (2022)
- Case-study seminar given in the International Summer School on Advanced Methods in Global Health course organized by ISGlobal (September 2021, Online)
- Epimutacions workshop given in the "Técnicas ómicas en el dignóstico de enfermedades raras" course organized by ISGlobal and CNAG (16 and 17 November 2022, Parc Científic de Barcelona)
- ISGlobal PhD representative (2021-2022)
- Bioinformatic Research Group in Epidemiology (BRGE) community manager (2020-2022)

Scientific divulgation tasks

- Nature Portfolio Post titled "The early-life exposome modulates the effect of polymorphic inversions on DNA methylation" - https://go.nature.com/3vUHwqO

- ISGlobal website post titled "Both Genetics and Environment Play a Role in DNA Methylation and Thereby in the Expression of Certain Genes" written by Adelaida Sarukhan - https://www.isglobal.org/en/-/tanto-la-genetica-como-el-ambiente-juegan-un-papel-en-la-metilacion-del-adn-y-por-lo-tanto-en-la-expresion-de-ciertos-genes

- El·lipse website post titled "What do your cells think about you smoking marijuana?" - https://ellipse.prbb.org/what-do-your-cells-think-about-you-smoking-marijuana/

- 100tífiques volunteer in 2021 and 2022. It is an initiative organized by the Barcelona Institute of Science and Technology (BIST) and Fundació catalana per la recerca i la innovació (FCRI) that aims to foment science among girls in secondary school.

- Air Pollution workshop for general audience during the PRBB science festival 2021

- Career guidance in the Biosciences Faculty 2022, UAB

- Q&A interview with TruDiagnostic

Presentations in congresses

- EUTOPIA Week in Ljubljana (21st to 25th November 2022) – Poster presentation. ONLINE.

  Title: The effect of polymorphic inversions on DNA methylation and its modulation by the early-life exposome.

- 8th ISGlobal PhD Symposium (5th October 2022) – Oral presentation. Barcelona, Spain.

  Title: What do your cells think about you smoking marijuana and drinking alcohol?

- Women in Data Science 2022 (28th September 2022) – Poster presentation. Barcelona, Spain.

  Title: The effect of polymorphic inversions on DNA methylation and its modulation by the early-life exposome.

- European Human Genetics Conference 2022 (11th to 14th June 2022) - Poster presentation. Vienna, Austria.

  Title: The effect of polymorphic inversions on DNA methylation and its modulation by the early-life exposome.

- Rin 4' UPF competition (24th May 2022) – Oral presentation. Barcelona, Spain.

  Title: ¿Qué piensan tus células de que fumes marihuana?

- 17th INMA Scientific Conferences (23rd to 24th November 2021) – Oral presentation. San Sebastián, Spain.

  Title: The effect of polymorphic inversions on DNA methylation and its modulation by the early-life exposome.

- 7th ISGlobal PhD Symposium (30th September 2021) – Oral presentation. Barcelona, Spain.

  Title: The state of the Y chromosome as a biomarker to monitor cancer in men.

- LMB-IBMB Graduate Life Sciences Symposium 2021 (30th June to 2nd July 2021) – Oral presentation. ONLINE

  Title: The state of Y chromosome as a biomarker to monitor cancer in men.

Courses

PRBB intervals courses:

- "Scientific writing: putting the why before the how"
- "Introduction to scientific publishing and how to read a paper"
- "Behind the scenes - a small group tutorial in oral presentation skills for scientists"
- "Gestión del estrés para profesionales de investigación"

- "Understanding career opportunities"
- "Business Opportunities in science and beyond"

BIST courses:

- "Writing thesis bootcamp"
- "Social Media"
- "CVs and Employability"

ISGlobal courses:

- "De las microagresiones al acoso. Las mujeres en investigación científica"
- "Concienciación de ciberseguridad"

DataSHIELD courses:

- "DataSHIELD Beginners Workshop"
- "DataSHIELD Resources"
- "DataSHIELD Statistical Analysis"
- "DataSHIELD Omics"

Other courses:

- "Data Visualization for Environmental Epidemiology with ggplot2" (US EPA)
- "Preparación y defensa de un póster científico" (Universidad de Granada)
- "Introduction to exploring genome-phenome data with EGA" (EMBL-EBI)
- "Podcasting" (LMB-IBMB)
- "Enhance your Job application skills: How to write your CV and tips to prepare a job interview" (UPF)
- "Técnicas ómicas en el dignóstico de enfermedades raras" (ISGlobal and CNAG)

9.2. Pre-processing TruDiagnostic data

General information

Array: IlluminaHumanMethylationEPIC
Quality control software: *meffil*
N (initial): 5,816 → N (final): 3,424
Probes (initial): 865,859 → Probes (end): 740,023

SampleSheet from idat files

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/A_Create_SampleSheet.R
  Time ~ 2 min

We first separated the initial idat files (6,187) in two folders:

- **idat_remove** (387 idat files)
  This folder contains the negligible data (most of them are fictitious data). It consists of 6 smaller folders: 08142020 Pre Open DMAPs, Duplicates, 205735180078, iScan Comp Comparison, Redo, and Test Run 02162022. We avoided these idat files.
- **idat_use** (5,816 idat files)
  Here, we collected all the idat files that should be considered in the following steps.

We created a SampleSheet using the idat files from the idat_use folder with the *meffil.create.sample.sheet* function. This function generated a *data.frame* of 5,816 rows (IDs).

Since we needed the sex annotation of these individuals for the sample Quality Control (QC), we compared the IDs from the SampleSheet with the ones in the Patient Metadata file (after removing those individuals with a BMI out of the range 10-60 and with intersex sex).

At the end, we obtained 3,599 individuals matching between the SampleSheet and the Metadata with Male/Female sex. However, we removed three individuals (205772280052_R06C01, 205772280052_R07C01, and 205772280052_R08C01) that showed troubles in the QC step and 6 individuals that were duplicated

(205772280137_R08C01, 205772280146_R01C01, 205772290045_R01C01, 205828610080_R05C01, 205832310130_R08C01, and 205832310143_R07C01), leading to a final SampleSheet of 3,590 individuals with biological sex annotated (**Figure 1**).

Sample Quality Control

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/B_Sample_QC.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/B_call_SampleQC.sh
  Time ~ 4h 30 min / Mem ~ 23 Gb

Using the SampleSheet previously mentioned with 3,590 IDs, we performed the sample QC using the *meffil.qc* function with the "blood gse35069 complete" as reference.

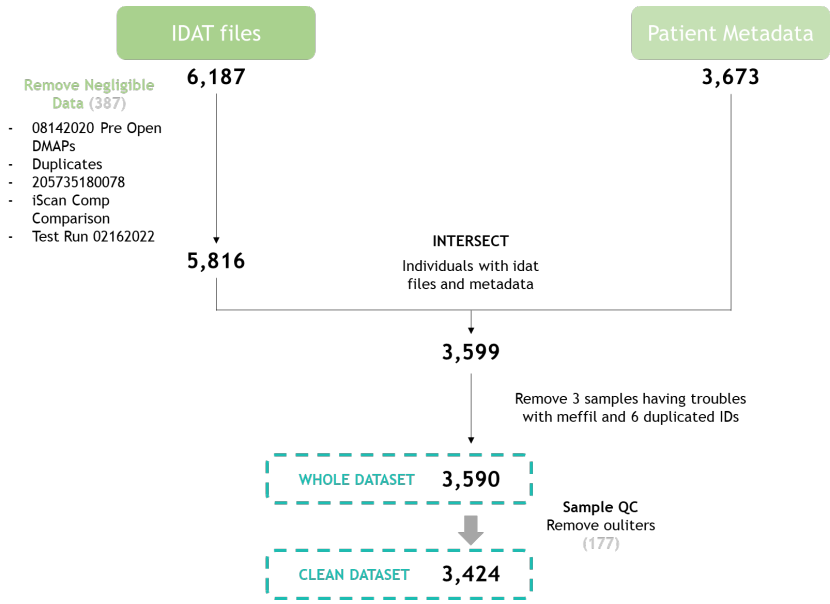


Figure 1 | Selection of the individuals based on the available metadata and the sample Quality Control (QC).

We used the default parameters for the QC report:

- detection.threshold = 0.01,
- bead.threshold = 3
- beadnum.samples.threshold = 0.05
- detectionp.samples.threshold = 0.05
- detectionp.cpgs.threshold = 0.05
- beadnum.cpgs.threshold = 0.05
- sex.outlier.sd = 3

The QC report file can be found at the following path:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/QC/qc-report_whole.html

The report showed outliers for a lot of conditions, but we only selected the outliers based on:

- Control probe (dye.bias) – 7
- Methylated vs Unmethylated - 74
- X-Y ratio outlier - 55
- Low bead numbers - 1
- Detection *P*-value - 1
- Sex mismatch - 36
- Control probe (bisulfite1) - 0
- Control probe (bisulfite2) - 0

Among them, we found 7 samples with more than one issue:

- 205676380102_R02C01
    - ❖ Sex mismatch
    - ❖ X-Y Ratio Outlier
- 205676390016_R08C01
    - ❖ Sex mismatch
    - ❖ Detection *P*-value
    - ❖ X-Y Ratio Outlier
- 205676390106_R03C01
    - ❖ Methylated vs Unmethylated
    - ❖ X-Y Ratio Outlier
- 205772280075_R02C01
    - ❖ Methylated vs Unmethylated

    - ❖ X-Y Ratio Outlier
- 205772280075_R03C01
    - ❖ Methylated vs Unmethylated
    - ❖ X-Y Ratio Outlier
- 205772280091_R04C01
    - ❖ Sex mismatch
    - ❖ X-Y Ratio Outlier
- 205832330027_R08C01
    - ❖ Methylated vs Unmethylated
    - ❖ X-Y Ratio Outlier

In total, the outliers represented 166 samples. We decided to remove all of them and continue the analysis with 3,424 individuals (**Figure 1**). We estimated the cellular composition based on methylation levels and we generated another QC report with these selected individuals:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/QC/qc-report_clean.html

Functional Normalization

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/C_Functional_Normalization.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/C_call_FunctNorm.sh
  Time ~ 11h 30min / Mem ~ 100 Gb

The next step is to normalize the CpG methylation values. To this end, we first estimate the number of principal components to use based on the methylation levels of the control probes (**Figure 2**). Looking at the plot, we considered that 10 PCs was a good approximation for the normalization.

Previously to generate the beta values, we set poorly detected methylation values to missing. Poor signal was identified during QC as signal that failed to pass the detection *P*-value threshold (0.01) or bead threshold (3). Moreover, we removed probes that have more than 5% of poorly detected values. In total, we removed 28,117 probes. Among them, 3,297 had poor detection *P*-value, 24,365 failed the bead threshold, and 455 failed both thresholds (**Figure 3**).

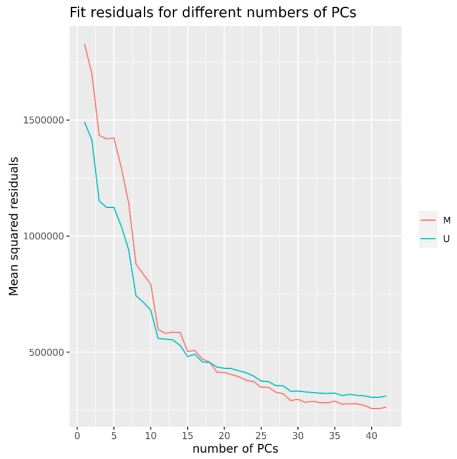


Figure 2 | Plot showing the fit of the residuals for different numbers of principal components (PCs). M: methylated; U: unmethylated.

During the normalization process, we decided to save the output to a GDS (Genomic Data Structure) because of the magnitude of the output and the high memory demand.

Once we got the normalized betas, we removed the CpG sites that had accumulated more than 5% of missing (1,110 probes), leading to a total of 836,632 CpG sites. In the case of the IDs, there were not individuals with more than 5% of missing. In the final norm.beta object we had 0.34% of missing.

Finally, we calculated the principal components of normalized betas based on the 50,000 most variable CpG sites (this is the value by default).

We created a normalization report using 4 variables as batches (slide, sex, Sentrix_Row, and Sentrix_Col) and the default parameters:

- control.pcs=1:5
- batch.pcs=1:5
- batch.threshold=0.01

The report can be accessible at:

/PROJECTES/GENOMICS/TruDiagnostic/prepro_files/Functional_Normalization/normalization-report_clean.html

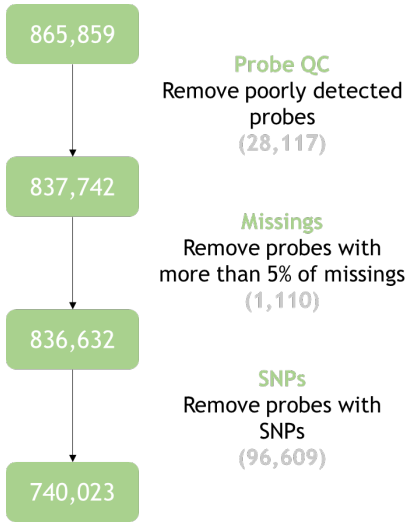


Figure 3 | Selection of the CpG sites based on the Quality Control (QC), the number of missing, and the probes with SNPs. The removal of probes with SNPs is based on the InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation.

Create GenomicRatioSet

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/D_Create_GenomicRatioSet.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/D_call_GRset.sh
  Time ~ 2h 15min / Mem ~ 185Gb

The GenomicRatioSet (GRset) is an object that can group different type of data:

- Beta values mapped to a genomic location
- Metadata → accessible with the pData() function
- Annotation data → accessible with the getAnnotation() function or with the rowData() in case you have included it there

We created a GRSet for our data using the normalized beta values from the norm.beta_clean object. For the metadata, we recodified some variables to simplify the further analyses (See "Descriptive_Analysis_metadata.html"). We joined the 120 metadata variable, the estimated cellular composition (7

variables), and the Slide variable and we included it in the GRset metadata. For the annotation information, we used the default EPIC annotation by Illumina (ilm10b4.hg19).

We created the GRSet object using the *makeGenomicRatioSetFromMatrix* function from the *minfi* package.

After creating the GRset object, we used InfiniumAnnotation from https://zwdzwd.github.io/InfiniumAnnotation to filter probes where 30bp 3'-subsequence of the probe is non-unique, probes with INDELs, probes with extension base inconsistent with specified color channel (type-I) or CpG (type-II) based on mapping, probes with a SNP in the extension base that causes a color channel switch from the official annotation, and probes where 5bp 3'-subsequence overlap with any of the SNPs with global population frequency higher than 1%.

Finally, we checked the last version in HGNC of the gene symbols and we changed the ones that were annotated using previous versions.

Our final GenomicRatioSet object contains:

- 3,424 columns (IDs)
- 740,423 rows (CpG sites)
- 128 columns in the colData (metadata + cellular composition + Slide)
- 6 columns in the rowData
- 46 columns in the annotation information

This object can be found at:

/PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_clean.Rdata

Principal Component Analysis

In the normalization-report, we observed that there was a slightly division of the normalized betas into two clusters when comparing PC1 and PC2 (**Figure 4**). We tested whether the first 3 PCs were associated with any of the variables from the metadata that were potential variables to show big differences in methylation: sex, ethnicity, slide, age, cell type and collection date (**Table 2**). In **Figure 4**, we can see graphically the association between

sex, ethnicity, and collection year with the first 3 PCs. In **Table 2**, we can see the most significant associations with their effect and significance. Among these, the different cell types, the slide, and the collection date are very associated with the first 3 components. The age, sex, and ethnicity are also associated with the PC1, but the significance and the correlation are lower.

Although we have found different variables that are associated with the PCs, we performed a Surrogate Variable Analysis (SVA) to detect batch variables that were unknown. To this end, we first imputed missing data.

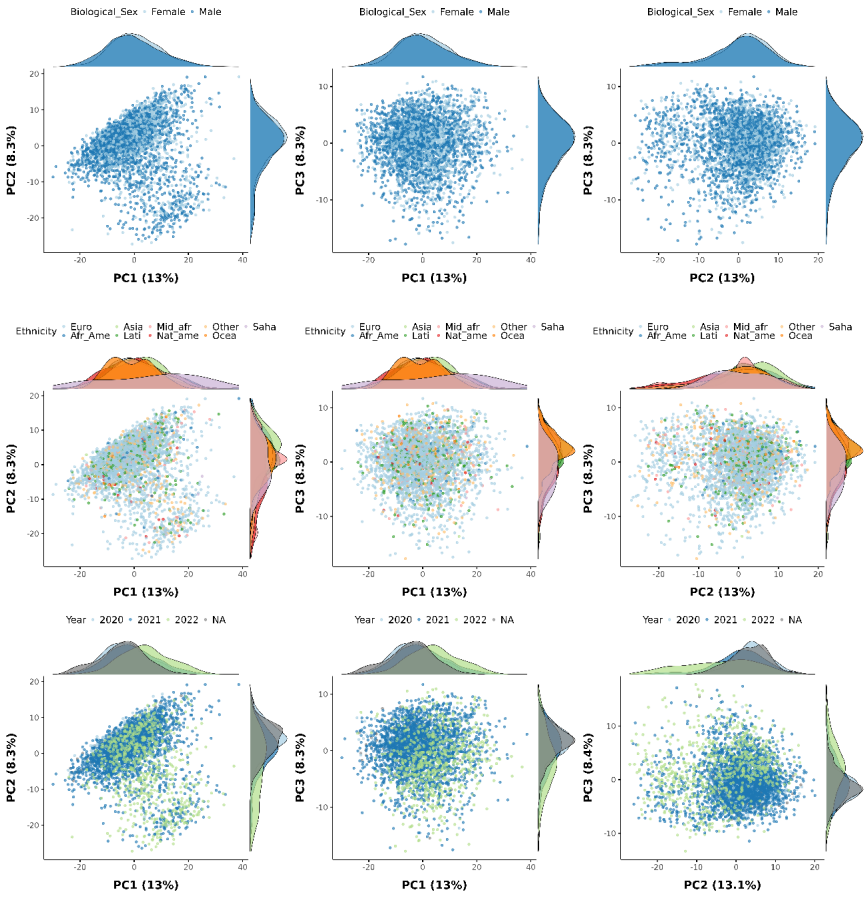


Figure 4 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with Biological sex, ethnicity, and collection year in normalized beta.

Table 2 | Significant pairwise associations (p<0.000001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in normalized beta. "Year" means the year when the sample was collected. "Year_Month" is calculated by multiplying the year by 12 plus the month.

x	y	test	p.value	estimate	r2
CD4T	PC1	F-test	0	2721.623083	0.442999684
Neu	PC1	F-test	0	10322.13186	0.751021015
NK	PC1	F-test	0	2323.246347	0.404377151
Bcell	PC1	F-test	1.1581E-273	1508.132326	0.305900983
Neu	PC2	F-test	1.0991E-138	689.9388481	0.16778918
CD4T	PC2	F-test	1.7935E-121	595.9287929	0.14831741
CD8T	PC2	F-test	1.43432E-94	453.3301882	0.116978468
CD8T	PC1	F-test	2.82849E-91	436.2226842	0.113063117
Year_Month	PC2	F-test	4.44468E-86	410.1926453	0.1102315
Slide	PC2	F-test	3.60693E-80	379.1273236	0.099740759
Year	PC2	F-test	1.21324E-69	326.3142668	0.08971297
Year_Month	PC1	F-test	4.74343E-52	238.7017295	0.067245574
Slide	PC1	F-test	2.0153E-48	220.6752853	0.060580554
NK	PC3	F-test	4.18784E-41	185.1520449	0.051329149
Year	PC1	F-test	5.52047E-40	179.9051618	0.051535391
CD8T	PC3	F-test	2.68267E-36	162.0260174	0.045207824
Year	PC3	F-test	3.12093E-36	161.837986	0.046601076
Year_Month	PC3	F-test	7.38517E-34	150.4612995	0.043467567
Mono	PC3	F-test	1.41682E-28	125.2231654	0.035301744
Slide	PC3	F-test	3.31958E-26	114.0206628	0.032245474
NK	PC2	F-test	1.78607E-19	82.45216775	0.023527834
Mono	PC2	F-test	1.03948E-18	78.88967139	0.022534178
Bcell	PC3	F-test	1.53303E-15	64.19268692	0.018413408
Biological_Sex	PC2	t-test	1.90839E-12	-1.82941353	0.014601133
Bcell	PC2	F-test	7.12737E-11	42.75478946	0.012339918
age	PC1	F-test	1.29471E-10	41.57303771	0.012002934
Mono	PC1	F-test	2.37329E-09	35.83178487	0.010362501
Ethnicity	PC2	t-test	3.44457E-07	-1.552481302	0.007647631

Impute data

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/F_Impute_Data.R
  Time ~ 25 min / Mem ~ 80Gb

To perform SVA, we need data without missing. Then, we have created another GRset with imputed betas based on the median of each CpG site:

/PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_clean_imp.Rdata

Again, we have tested pairwise associations between the previous variables and the first 3 PCs and the results are very similar compared with the non-imputed normalized betas (**Figure 5** and **Table 3**). Therefore, we can assume that the imputation is not altering our data.

Surrogate Variable Analysis

- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/G_SVA.R
- /PROJECTES/GENOMICS/TruDiagnostic/Processing_Scripts/Meffil/G_call_SVA.sh
  Time ~ 1 day / Mem ~ 150Gb

In the SVA analysis, we included some covariates in the model in order to keep their effect on DNA methylation: marijuana, biological sex, ethnicity, age, neuropsychological, cardiovascular, respiratory, and endocrine diseases, tobacco, alcohol, amphetamines, benzodiazepines, hallucinogens, and MDMA use, and drug or alcohol addiction for mother or father. We did not include cell type nor slide nor collection date because they were not variables of our interest, and we want to remove their effect on DNA methylation.

First, we estimated the number of surrogate variables (SVs) using isva::EstDimRMT and it was 127. Since it was a huge number of SVs, we decided to follow the guidelines from GTEX where they recommend using 60 SVs when N>350. (https://gtexportal.org/home/documentationPage#staticTextAnalysisMethods).

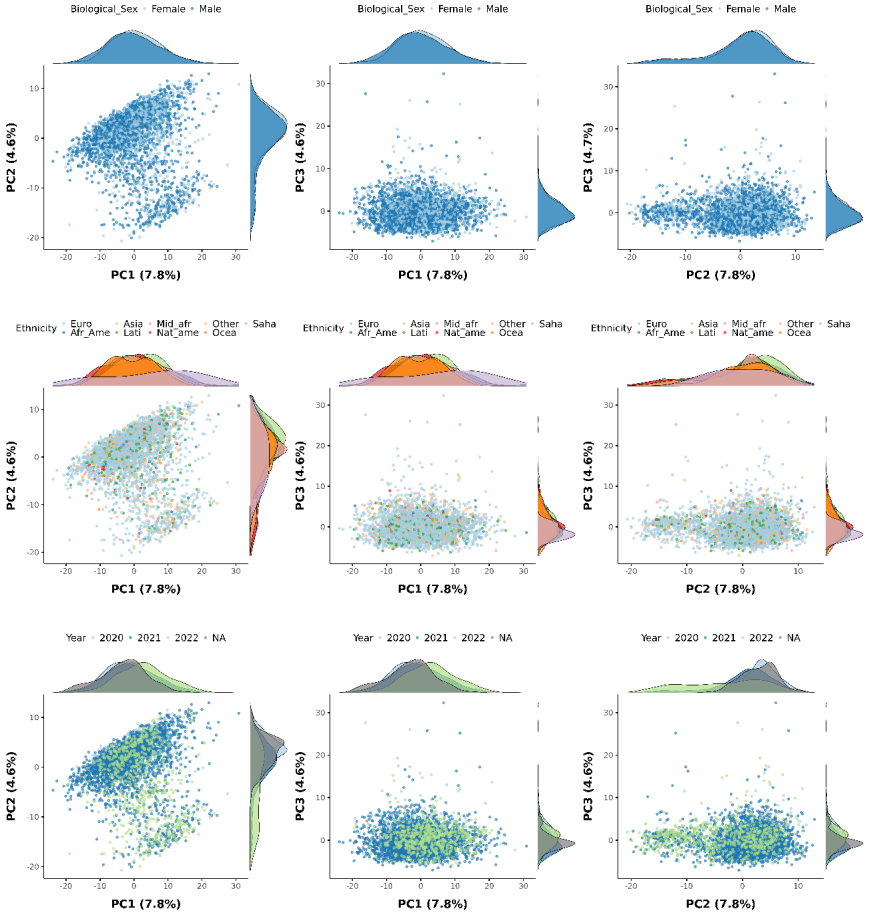


Figure 5 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with sex, ethnicity, and collection date in imputed normalized beta.

Second, we calculated the 60 SVs from our data, and we saved the object:

- /PROJECTES/GENOMICS/TruDiagnostic/prepro_files/batch/sv.obj.Rdata

Third, we adjusted the beta values by these SVs, and we created a new GRset with the residuals:

- /PROJECTES/GENOMICS/TruDiagnostic/Final_datasets/GRset_SVA.Rdata

Table 3 | Significant pairwise associations (p<0.000001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in imputed normalized beta. "Year" means the year when the sample was collected. "Year_Month" is calculated by multiplying the year by 12 plus the month.

x	y	test	p.value	estimate	r2
CD4T	PC1	F-test	0	3298.833	0.490837
Neu	PC1	F-test	0	14642.29	0.810565
NK	PC1	F-test	0	2408.706	0.413107
Bcell	PC1	F-test	1.6E-276	1527.071	0.308557
CD8T	PC1	F-test	9.3E-113	549.1642	0.138288
Year_Month	PC2	F-test	4.7E-89	425.5886	0.113898
Neu	PC2	F-test	4.92E-85	404.0305	0.1056
Slide	PC2	F-test	1.22E-84	402.0029	0.105126
CD4T	PC2	F-test	9.79E-79	371.8198	0.098007
Year	PC2	F-test	7.02E-73	342.6801	0.09379
CD8T	PC2	F-test	7.93E-66	306.6752	0.082248
Year_Month	PC1	F-test	8.34E-36	159.7895	0.046038
Mono	PC3	F-test	1.62E-35	158.2805	0.044209
Slide	PC1	F-test	2.67E-33	147.6898	0.041373
Year	PC1	F-test	2.06E-27	119.7947	0.034917
CD8T	PC3	F-test	7.9E-19	79.44316	0.022689
Biological_Sex	PC2	t-test	1.2E-18	-1.7044	0.02303
NK	PC3	F-test	1.2E-17	73.95152	0.021153
Mono	PC2	F-test	1.48E-12	50.44953	0.014529
Mono	PC1	F-test	2.82E-12	49.16834	0.014165
age	PC1	F-test	9.3E-12	46.79456	0.01349
Ethnicity	PC2	t-test	3.58E-10	-1.44119	0.011619
NK	PC2	F-test	5.33E-10	38.77541	0.011204
Ethnicity	PC2	t-test	3.05E-08	-1.52286	0.009507
Biological_Sex	PC3	t-test	9.09E-08	0.535045	0.008388

Finally, we calculated the first PCs from these residuals, and we evaluated again the association of these PCs with the previously variables mentioned (**Figure 6** and **Table 4**).

Table 4 | Significant pairwise associations (p<0.00001) between variables (sex, ethnicity, slide, age, cell type, and collection date) and the first 3 Principal Components (PCs) in SVA.

x	l	y	test	p.value	estimate	r2
Ethnicity		PC2	F-test	1.9412E-220	152.7985525	0.263594142
Biological_Sex		PC1	F-test	1.0874E-124	613.3341816	0.151990927
Biological_Sex	Male	PC1	t-test	2.9545E-123	-0.922146543	0.151087284
Biological_Sex	Female	PC1	t-test	9.3239E-123	0.91750348	0.150596286
Ethnicity	Other	PC2	t-test	1.8096E-114	1.285252055	0.142778418
Ethnicity	Ocea	PC2	t-test	5.2228E-100	3.736507287	0.127807482
Ethnicity	Asia	PC2	t-test	5.43227E-70	2.313987409	0.090544531
Ethnicity	Euro	PC2	t-test	8.12027E-52	-0.53141297	0.067179053
Ethnicity	Afr_Ame	PC3	t-test	1.19913E-51	-1.741490472	0.064962381
Ethnicity		PC3	F-test	8.30806E-49	31.95998289	0.069654634
Biological_Sex	Male	PC2	t-test	1.26926E-47	0.489724572	0.060903123
Ethnicity	Afr_Ame	PC2	t-test	3.12632E-25	1.003281787	0.0323071
Ethnicity		PC1	F-test	1.36925E-18	12.98417007	0.029518925
Ethnicity	Afr_Ame	PC1	t-test	3.32232E-17	-1.160572825	0.020770515
Biological_Sex		PC2	F-test	7.53485E-17	70.24919389	0.020115745
Ethnicity	Lati	PC2	t-test	3.56142E-14	0.383422047	0.017339014
Biological_Sex	Female	PC2	t-test	1.53287E-09	-0.207886878	0.010819591
Ethnicity	Euro	PC3	t-test	4.7569E-09	0.228699768	0.009996995

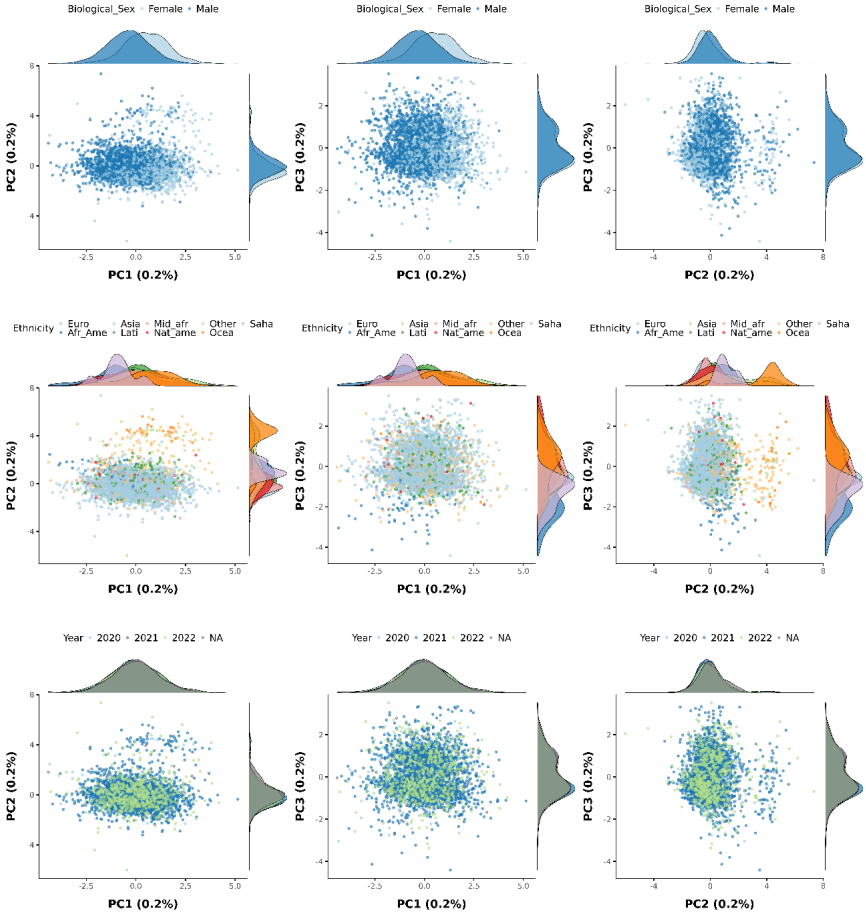


Figure 6 | Principal Component Analysis comparing the first 3 Principal Components (PCs) with biological sex, ethnicity, and collection date in SVA residuals.

We can see that the associations with cell type, slide, and collection date have disappeared, most likely because the SVs estimated explain most of their variability. To prove this, we have tested correlation between SVs and these variables (**Figure 7**). First, it is worth mentioning that slide is highly associated with Year_Month variable (r^2=0.837). This did not surprise us because the different slides have been used in different days or months. In addition, the different cell types are also correlated. Second, we can see that SV1 and SV2 are the two surrogate variables that are more correlated with covariates.

To see these correlations more in detail, we evaluated the pair-wise associations between a lot of variables (including drugs consumption and some diseases) with SVs (**Table 5**). Among them, we can see again slide, cell type, and collection date altogether with sex and age. Sex is mainly correlated with SV19 and age is mainly correlated with SV17.

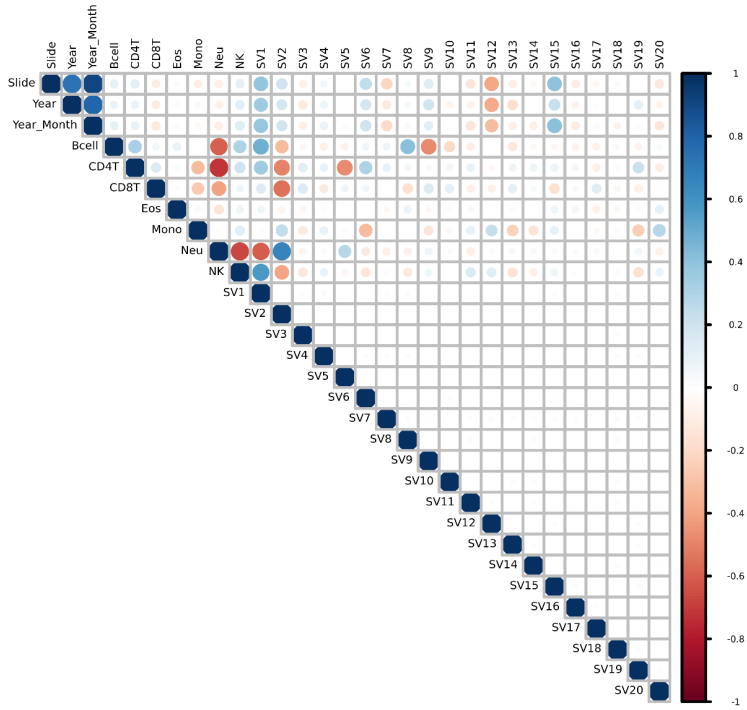
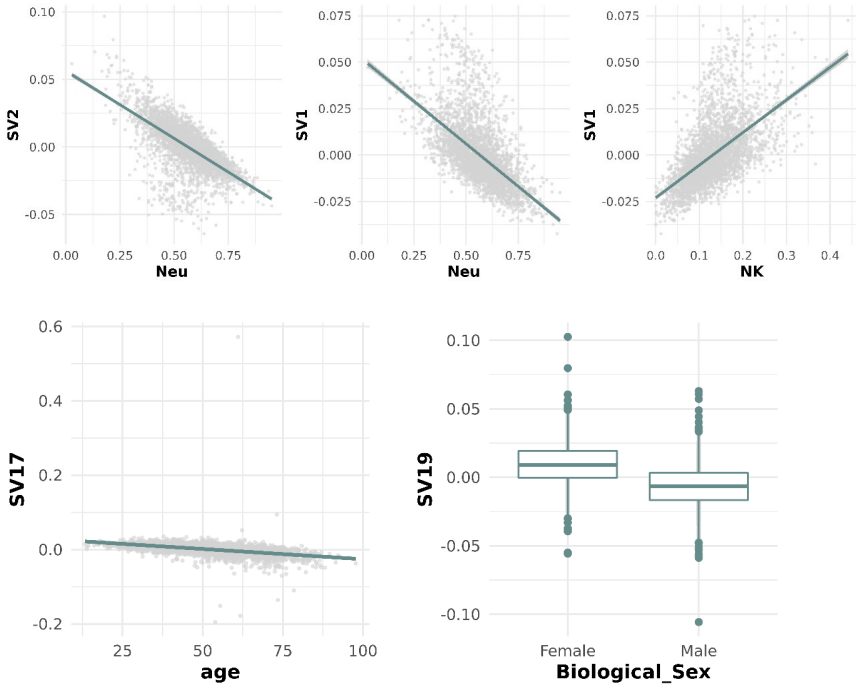


Figure 7 | Correlation plot between slide, collection date (Year and Year_Month), cell types (Bcell, CD4T, CD8T, Eos, Mono, Neu, and NK) with the first 20 surrogate variables (SVs).

Table 5 | Significant pairwise associations (r^2<0.1) between variables (sex, ethnicity, slide, age, cell type, collection date, drugs, alcohol, and neurological, cardiovascular, respiratory, and endocrine diseases) and surrogate variables (SVs).

x	l	y	test	p.value	estimate	r2
Neu		SV2	F-test	0	2808.473	0.450764031
Neu		SV1	F-test	0	1971.284	0.365507137
NK		SV1	F-test	0	1751.664	0.338573201
CD8T		SV2	F-test	2.8E-269	1479.149	0.30179633
Bcell		SV1	F-test	5.4E-212	1115.784	0.245887396
CD4T		SV2	F-test	2.3E-211	1111.932	0.245246688
CD4T		SV5	F-test	8.6E-195	1012.049	0.228244849
Bcell		SV9	F-test	7E-186	959.2464	0.218943723
Biological_Sex	Female	SV19	t-test	6E-176		0.209365268
Biological_Sex		SV19	F-test	5.6E-171	872.3528	0.203139519
Biological_Sex	Male	SV19	t-test	9.7E-170		0.202708486
age		SV17	F-test	2.7E-145	726.6454	0.175152449
Bcell		SV8	F-test	1.5E-143	716.8678	0.173203836
Slide		SV15	F-test	2E-140	699.5484	0.169729512
Slide		SV1	F-test	5.3E-140	697.2175	0.169259686
Year_Month		SV15	F-test	1.2E-134	669.6864	0.168233907
NK		SV2	F-test	1E-127	629.8141	0.155440029
Year_Month		SV1	F-test	1.6E-119	587.0054	0.150591228
Year		SV12	F-test	6.4E-114	556.7952	0.14395674
Slide		SV12	F-test	3.6E-117	572.7905	0.143384361
Year		SV1	F-test	4.1E-100	483.3929	0.12739663
CD4T		SV1	F-test	4.9E-101	487.1228	0.124611782
age		SV16	F-test	5.6E-88	419.1817	0.109128321
Bcell		SV2	F-test	1.57E-81	386.0765	0.10138361



Summary GRsets

	Description GenomicRatioSet
GRset_clean	Normalized beta with missing values after QC sample
GRset_clean_imp	Normalized beta with imputed values using median method
GRset_ SVA	Residuals after adjusting the normalized and imputed betas by the 60 SVs