bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Identification of blood autosomal cis-expression quantitative trait methylation (cis-eQTMs) in children

Carlos Ruiz-Arenas[1,2], Carles Hernandez-Ferrer[2,3,4], Marta Vives-Usano[2,4,5], Sergi Marí[2,4,6], Inés Quintela[7], Dan Mason[8], Solène Cadiou[9], Maribel Casas[2,4], Sandra Andrusaityte[10], Kristine Bjerve Gutzkow[11], Marina Vafeiadi[12], John Wright[8], Johanna Lepeule[9], Regina Grazuleviciene[10], Leda Chatzi[13], Ángel Carracedo[14,15], Xavier Estivill[16], Eulàlia Martí[17], Geòrgia Escaramís[17], Martine Vrijheid[2,4,6], Juan Ramon González[2,4,6]*, Mariona Bustamante[2,4,6]*

Affiliations:

1. Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Barcelona, Spain

2. Universitat Pompeu Fabra (UPF), Barcelona, Spain

3. Centro Nacional de Análisis Genómico (CNAG-CRG), Center for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona, Catalonia, Spain.

4. ISGlobal, Dr Aiguader 88, 08003 Barcelona, Spain

5. Center for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Av Aiguader 88, 08003 Barcelona, Spain

6. CIBER Epidemiología y Salud Pública (CIBERESP), Barcelona, Spain

7. Medicine Genomics Group, University of Santiago de Compostela, CEGEN-PRB3, 15782, Santiago de Compostela, Spain

8. Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, UK

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9. University Grenoble Alpes, Inserm, CNRS, Team of Environmental Epidemiology Applied to Reproduction and Respiratory Health, IAB, 38000 Grenoble, France

10. Department of Environmental Science, Vytautas Magnus University, 44248 Kaunas, Lithuania

11. Department of Environmental Health, Norwegian Institute of Public Health, Oslo, Norway

12. Department of Social Medicine, University of Crete, Greece

13. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, USA

14. Medicine Genomics Group, CIBERER, University of Santiago de Compostela, CEGEN-PRB3, 15782, Santiago de Compostela, Spain

15. Galician Foundation of Genomic Medicine, IDIS, SERGAS, 15706, Santiago de Compostela, Spain

16. Quantitative Genomics Medicine Laboratories (qGenomics), Esplugues del Llobregat, Barcelona, Catalonia, Spain.

17. Departament de Biomedicina, Institut de Neurociències, Universitat de Barcelona, Barcelona, Spain

**Corresponding authors:**

Carlos Ruiz-Arenas: carlos.ruiza@upf.edu

Mariona Bustamante: mariona.bustamante@isglobal.org

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Abstract

**Background:** The identification of expression quantitative trait methylation (eQTMs), defined as correlations between gene expression and DNA methylation levels, might help the biological interpretation of epigenome-wide association studies (EWAS). We aimed to identify autosomal cis-eQTMs in child blood, using data from 832 children of the Human Early Life Exposome (HELIX) project.

**Methods:** Blood DNA methylation and gene expression were measured with the Illumina 450K and the Affymetrix HTA v2 arrays, respectively. The relationship between methylation levels and expression of nearby genes (transcription start site (TSS) within a window of 1 Mb) was assessed by fitting 13.6 M linear regressions adjusting for sex, age, and cohort.

**Results:** We identified 63,831 autosomal cis-eQTMs, representing 35,228 unique CpGs and 11,071 unique transcript clusters (TCs, genes). 74.3% of these cis-eQTMs were located at <250 kb, 60.0% showed an inverse relationship and 23.9% had at least one genetic variant associated with the methylation and expression levels. They were enriched for active blood regulatory regions. Adjusting for cellular composition decreased the number of cis-eQTMs to 37.7%, suggesting that some of them were cell type-specific. The overlap of child blood cis-eQTMs with those described in adults was small, and child and adult shared cis-eQTMs tended to be proximal to the TSS, enriched for genetic variants and with lower cell type specificity. Only half of the cis-eQTMs could be captured through annotation to the closest gene.

**Conclusions:** This catalogue of blood autosomal cis-eQTMs in children can help the biological interpretation of EWAS findings, and is publicly available at https://helixomics.isglobal.org/.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


# Keywords

eQTM, quantitative trait, epigenetics, DNA methylation, transcription, gene expression, blood, children, EWAS


# Abbreviations

BivFlnx: flanking bivalent region

CpG: cytosine nucleotide followed by a guanine nucleotide

eQTM: expression quantitative trait methylation

eQTL: expression quantitative trait locus

Enh: enhancer

EnhBiv: bivalent enhancer

EnhG: genic enhancer

EWAS: epigenome-wide association study

FC: fold change

FDR: false discovery rate

GO: gene ontology

GWAS: genome-wide association study

HELIX: Human Early-Life Exposome project

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Het: heterochromatin

IQR: interquartile range

meQTL: methylation quantitative trait locus

OR: odds ratio

Quies: quiescent region

ReprPC: repressed Polycomb

ReprPCWk: weak repressed polycomb

SE: standard error

SNP: single nucleotide polymorphism

TC: transcript cluster

TSS: transcription start site

TssA: active transcription start site

TssAFlnk: flanking active transcription start site

TssBiv: bivalent transcription start site

TSS200: proximal promoter, from TSS to 200 bp

TSS1500: distal promoter, from 200 bp to 1,500 bp

Tx: transcription region

TxFlnk: transcription at 5’ and 3’

TxWk: weak transcription region

3’UTR: 3’ untranslated region

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5'UTR: 5' untranslated region

ZNF.Rpts: zinc finger genes and repeats

# Background

Cells from the same individual, although sharing the same genome sequence, differentiate into diverse lineages that finally give place to specific cell types with unique functions. This is orchestrated by the epigenome, which regulates gene expression in a cell/tissue and time-specific manner [1–3]. Besides the central role of the epigenome in regulating embryonic and fetal development, X-chromosome inactivation, genomic imprinting, and silencing of repetitive DNA elements, it is also responsible for the plasticity and cellular memory in response to environmental perturbations [1–3]. When this happens during prenatal or early life, the re-programmed epigenome might not match the later environment, and this can lead to increased disease risk. This is known as the Developmental Origins of Health and Disease (DoHAD)[4]. In addition to the plasticity or re-programming hypothesis, environmental insults experienced during prenatal life may directly disrupt the correct development of organs without a homeostatic response, which might also have long-term consequences on health [4].

Massive epigenetic alterations, caused by somatic mutations or as a result of age and/or injury, were initially described in cancer [3]. The paradigm of environmental factors modifying the epigenome and leading to increased disease risk was then extrapolated from cancer to a wide range of common diseases. As a consequence, a high number of epigenome-wide association studies (EWAS) investigating the association between prenatal and postnatal exposure to environmental factors and DNA methylation, and between DNA methylation and disease [1,3] have been performed in he recent years. EWAS have found thousands of associations between DNA methylation and environmental exposures or disease, which have been inventoried in two catalogues: the EWAS catalogue [5] and the EWAS Atlas [6].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

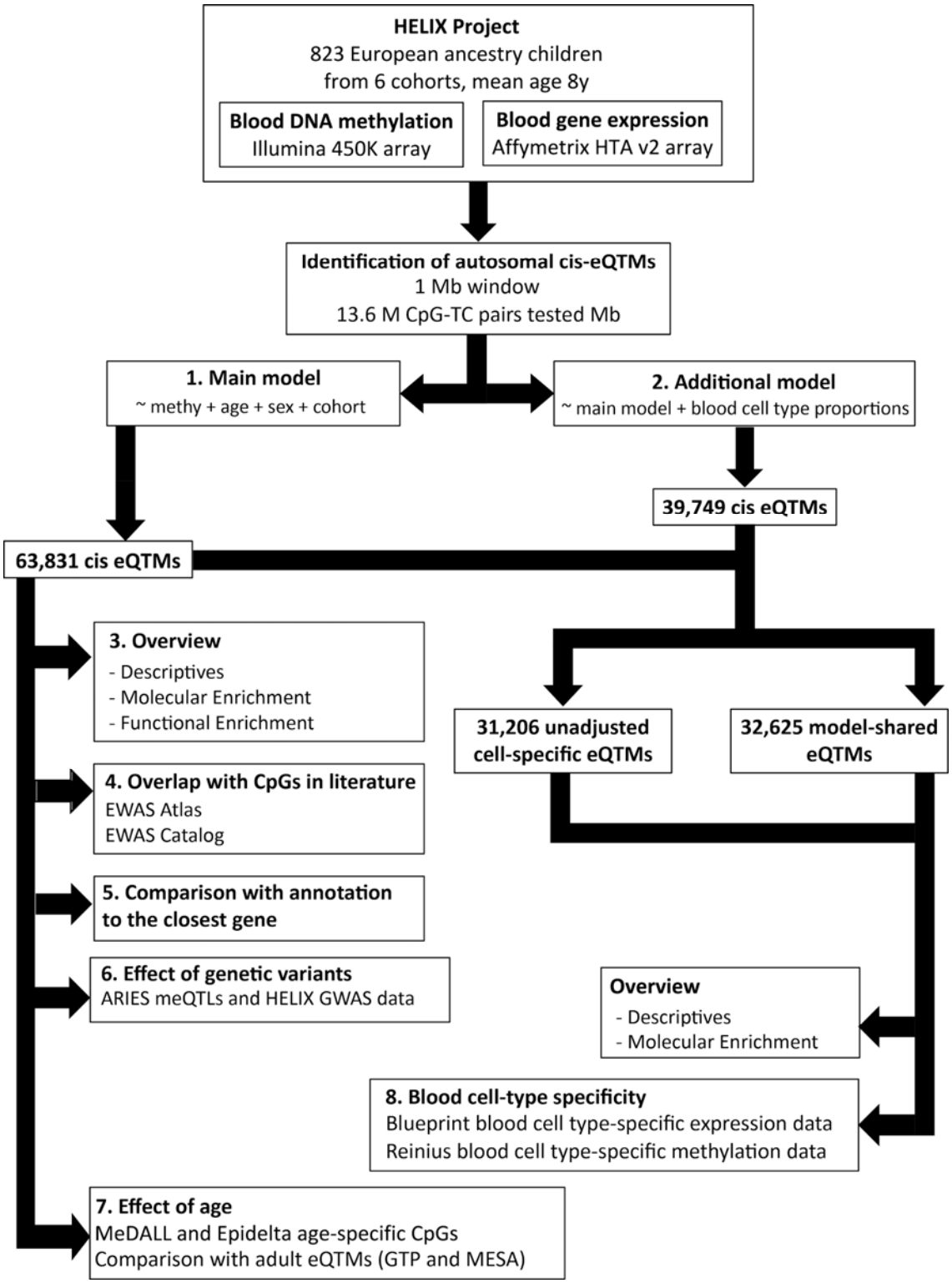
For instance, the latter includes 0.5 million associations for 498 traits from 1,216 studies, including 155 different cells/tissues.

Despite the success of EWAS in identifying altered methylation patterns, the role of genetic background, access to the target tissue/cell, confounding, reverse causation and biological interpretation are still challenging issues [1,3]. Regarding the latter, most studies do not dispose of transcriptional data to test the consequences of DNA methylation changes in gene expression. When gene expression is not available, a common approach is to assume that CpGs affect the expression of the closest gene [7]. Although this approach is easy to implement, it is limited in the fact that CpGs might regulate distant genes or might not regulate any gene at all [1,8]. Another approach to interpreting findings is to search for regulated genes of candidate CpGs in expression quantitative trait methylation (eQTM) studies, i.e. genome-wide studies investigating the associations between DNA methylation levels and gene expression [9,10]. Several eQTM studies have been published in diverse cell types/tissues: whole blood [8,11], monocytes [11–13], lymphoblastoid cell lines, T-cells and fibroblasts derived from umbilical cords [14,15], fibroblasts [16], liver [17], skeletal muscle [18], nasal airway epithelium [19] and placenta [20]. As most of the EWAS are conducted in whole blood [6,21], there is a need for comprehensive eQTM studies in this tissue. Available whole blood eQTM studies to date only cover samples from adults [8,11] and their validity in children has not been assessed. Besides, they do not consider the effect of genetics and blood cellular composition.

In this study, we analyzed DNA methylation and gene expression data from the Human Early-Life Exposome (HELIX) project to generate the first blood autosomal cis-eQTM catalogue in children (https://helixomics.isglobal.org/). We characterized child blood autosomal cis-eQTMs at the molecular level, compared them with eQTMs identified in adults, analyzed the proportion of cis-eQTM CpG-gene pairs captured through annotation to the closest gene, and assessed the influence of genetic variation and blood cell type specificity on the association between methylation and gene expression in eQTMs. An

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

overview of all the analyses can be found in Figure 1. This public resource will help the functional interpretation of EWAS findings in children.



bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Figure 1. Analysis workflow.** The figure summarizes the analyses conducted in this study. The first step was the identification of blood autosomal cis-eQTMs (1 Mb window) in 823 European ancestry children from the HELIX project by running two models: the main model adjusted for age, sex, and cohort (1); and an additional model further adjusted for blood cell type proportions (2). Then, we characterized cis-eQTMs identified in the main model by performing different enrichment analyses (3), analyzing their overlap with CpGs described in the literature (4), evaluating the proportion of cis-eQTM CpG-gene pairs captured through annotation to the closest gene (5), assessing the effect of genetic variants (6) and age by checking the overlap with eQTMs described in adults (7). Finally, we used cis-eQTMs identified in the additional model to investigate blood cell type specificity (8).

# Results

## Study population and molecular data

The study includes 823 children of European ancestry from the HELIX project with available DNA methylation and gene expression data. These children, enrolled in 6 cohorts, were aged between 6 and 11 years old and were sex balanced (Table 1).

**Table 1. Descriptive of the study population.**

| Variable | | N (%) |
|---|---|---|
| Cohort | BIB | 80 (9.7%) |
| | EDEN | 80 (9.7%) |
| | KANC | 143 (17.4%) |
| | MOBA | 188 (22.8%) |
| | RHEA | 154 (18.7%) |
| | SAB | 178 (21.6%) |
| Sex | Female | 372 (45.2%) |
| | Male | 451 (54.8%) |
| Variable | | Median (IQR) |

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

| | | |
|---|---|---|
| Age | | 8.06 (6.49-8.86) |
| Blood cell type proportions | Natural Killer | 0.02 (0.00-0.05) |
| | B-cell | 0.11 (0.9-0.14) |
| | CD4+ T-cell | 0.19 (0.15-0.23) |
| | CD8+ T-cell | 0.13 (0.10-0.16) |
| | Monocytes | 0.08 (0.07-0.10) |
| | Granulocytes | 0.44 (0.37-0.52) |

IQR: interquartile range

To initially explore methylation data, the 386,518 autosomal CpGs were classified according to their median methylation levels in low (0.0-0.3), medium (0.3-0.7) and high (0.7-1.0) [22]. Low (41.8%) and high (47.7%) methylation levels were the most abundant CpG categories. We also classified them as invariant (45.0%) or variant (55.0%) based on a methylation range threshold of 0.05 points, measured as the difference between the methylation values in percentile 1 and percentile 99 [23]. As expected, CpGs with medium methylation levels, which likely represent CpGs whose methylation status changes among blood cell types or which are influenced by genetic variants, showed higher variability in the population (p-value < 2e-16, Figure S1).

Gene expression data comprised 58,254 transcript clusters (TCs). Of those, 23,054 TCs encoded for a protein, according to the Affymetrix annotation. TCs are defined as groups of one or more probes covering a region of the genome, reflecting all the exonic transcription evidence known for the region, and corresponding to a known or putative gene.

We paired each TC to all CpGs closer than 500 Kb from its transcription start site (TSS), either upstream or downstream (1 Mb window around the TSS). In total, we obtained 13,615,882 TC-CpG pairs, 100 CpGs were not paired to any TC, and 189 TCs were not paired to any CpG, most of them being non-coding (92.1%). TCs were more promiscuous

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

than CpGs: each TC was paired to a median of 162 CpGs (interquartile range (IQR): 93; 297) while each CpG was paired to a median of 30 TCs (IQR: 20; 46) (Figure S2).

# Overview of blood autosomal cis-eQTMs in children

## Identification and classification of blood autosomal cis-eQTMs

We tested the association between DNA methylation and gene expression levels of the 13.6 million autosomal TC-CpG pairs through linear regressions. After correcting for multiple testing (see Material and Methods), we identified 63,831 significant child blood autosomal cis-eQTMs (0.47% of total TC-CpG pairs). For simplicity, we will refer to them as eQTMs in the subsequent text. These eQTMs comprised 35,228 unique CpGs and 11,071 unique TCs, of which 7,878 were annotated as coding genes. 38,310 eQTMs (60.0%) showed inverse associations, meaning that higher DNA methylation was associated with lower gene expression. Each TC was associated with a median of 2 CpGs (IQR = 1; 6), while each CpG was associated with a median of 1 TC (IQR = 1; 2) (Figure S3). As expected, CpGs in eQTMs were enriched for CpGs variable in the population (odds ratio (OR) = 6.1, p-value < 2.2e-16) (Figure S4) and for TCs with call rates >90% (OR = 5.4, p-value < 2.2e-16) (Figure S5). The complete catalogue of eQTMs can be downloaded from https://helixomics.isglobal.org/.

Then, we classified the CpGs in eQTMs into 5 types following two criteria: (1) the number of TCs affected, distinguishing between mono-CpGs (CpGs associated with a unique TC), and multi-CpGs (CpGs associated with two or more TCs); and (2) the direction of the effect, distinguishing between CpGs in inverse eQTMs, CpGs in positive eQTMs, and bivalent-CpGs (CpGs that exhibit inverse effects on some TCs and positive effects on others). Mono-CpGs in inverse eQTMs were the most abundant type (35.9%) (Table 2). CpG types were not independent: mono-CpGs were enriched for CpGs with positive effects compared to multi-CpGs (OR = 1.23, p-value < 2.2e-16).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**Table 2. Number of CpGs in child blood autosomal cis-eQTMs by type.**

| | CpGs in inverse eQTMs (N, %) | CpGs in positive eQTMs (N, %) | Bivalent-CpGs (N, %) | Total (N, %) |
|---|---|---|---|---|
| Mono-CpGs | 12,637 (35.9%) | 8,961 (25.4%) | 0, by definition | 21,598 (61.3%) |
| Multi-CpGs | 6,135 (17.4%) | 3,530 (10.0%) | 3,965 (11.3%) | 13,630 (38.7%) |
| Total | 18,772 (53.3%) | 12,491 (35.4%) | 3,965 (11.3%) | 35,228 (100%) |

Table percentages refer to the total number of CpGs in autosomal cis-eQTMs.

## CpGs in eQTMs are close to their target TC and have modest effects

CpGs whose methylation level was associated with gene expression tended to be close to the TSS, being 74.3% of all eQTMs located at <250 kb (Figure 2). Globally, the median distance between a CpG and a TC TSS in an eQTM was 1.3 kb (IQR = -84 kb; 117 kb). The observed downstream shift can be explained because for each TC we chose the most upstream TSS, which might not represent the TSS that gives the most abundant transcript in the blood (Figure 2). This shift depended on the direction of the effect: median distance for positive and inverse eQTMs were 6.7 kb (IQR = -79 kb; 103 kb) and 0.8 kb (IQR = -93 kb; 136 kb), respectively. A similar shift was observed for eQTLs (expression quantitative trait loci, i.e. single nucleotide polymorphisms (SNPs) associated with gene expression) in the Genotype-Tissue Expression (GTEx) project [24].

We report the effect size of eQTMs as the $\log_2$ fold change (FC) of gene expression per 0.1 points increase in methylation (or 10 percentile increase). In absolute terms, the median effect size was 0.12, being the minimum 0.002 and the maximum 16.4, and with 91.4% of the eQTMs with an effect size <0.5. A median effect size of 0.12 means that a change of 0.1 points in methylation levels was associated with around a 9% increase/decrease of gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression. We did not find any association between the effect size and the CpG-TC TSS distance or the relative position to the TSS (upstream or downstream) (Figure S6).

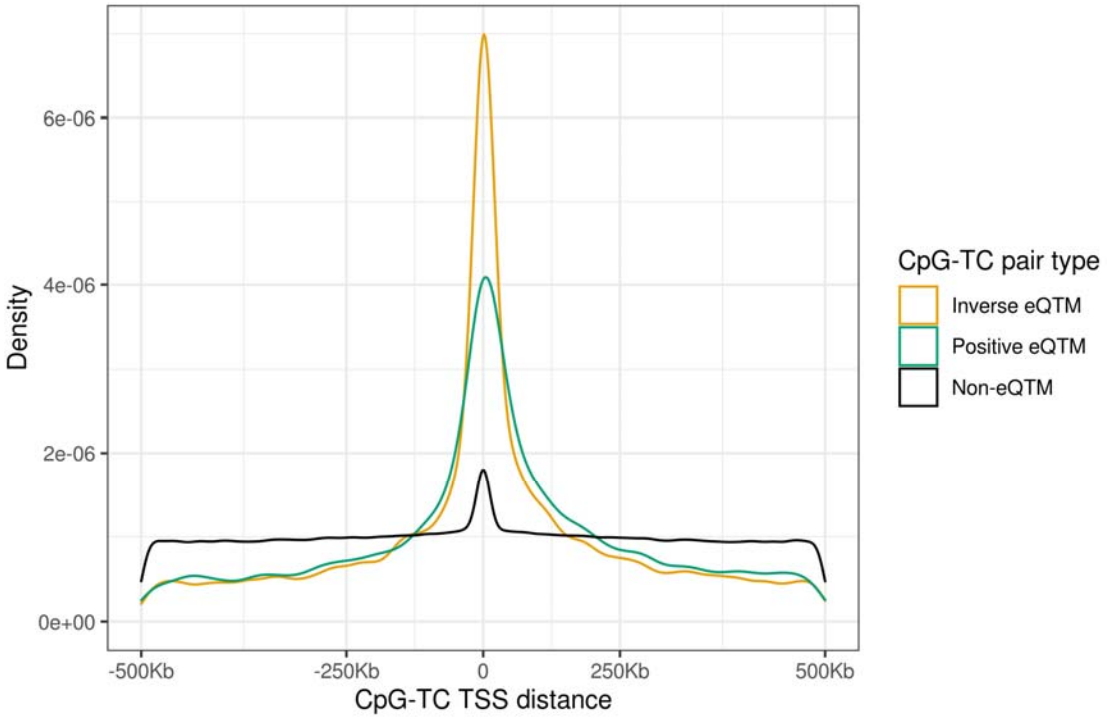



**Figure 2. Distribution of the distance between CpG and TC TSS by type of CpG-TC pair.** CpG-TC pairs were classified in: non-eQTMs (CpGs not belonging to any eQTM, in black); inverse eQTMs (inversely associated CpG-TC pairs defined as eQTMs, in yellow); and positive eQTMs (positively associated CpG-TC pairs defined as eQTMs, in green). Distance between CpG and TC TSS is expressed in kb.

## CpGs in eQTMs are enriched for blood active chromatin states and medium methylation levels

We characterized CpGs in eQTMs by evaluating their enrichment for diverse regulatory elements, including CpG island relative position and 15 chromatin states retrieved from 27 blood cell types from the ROADMAP Epigenomics project [25]. CpGs in eQTMs were depleted for CpG islands ($OR_{CpG\text{-}Island}$ = 0.51, $p\text{-value}_{CpG\text{-}Island}$ < 2e-16) while enriched for all the other positions (CpG island shores and shelves and open sea) (Figure 3A). CpGs in positive and inverse eQTMs showed quite similar distributions. Details on the proportion of eQTMs in each CpG island relative position can be seen in Figure S7A.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

CpGs in eQTMs were enriched for several active blood chromatin states: flanking active TSSs ($OR_{TssAFlnk}$ = 1.38, p-value$_{TssAFlnk}$ = 1.38e-181), weak transcription regions ($OR_{TxWk}$ = 1.54, p-value$_{TxWk}$ < 2e-259), enhancers ($OR_{Enh}$ = 2.43, p-value$_{Enh}$ < 2e-259), genic enhancers ($OR_{EnhG}$ = 1.82, p-value$_{EnhG}$ = 1.1e-259), and zinc finger genes and repeats ($OR_{ZNF.Rpts}$ = 1.7, p-value$_{ZNF.Rpts}$ = 9.48e-58) (Figure 3B; Figure S7B for proportions). In these regions, CpGs in positive and inverse eQTMs presented similar enrichments, except for flanking active TSSs (TssAFlnk) and genic enhancers (EnhG) for which enrichment was specific to CpGs in inverse eQTMs. Moreover, CpGs in positive eQTMs were specifically depleted for active transcription start sites ($OR_{TssA}$ = 0.58, p-value$_{TssA}$ = 1.22e-149) and transcription at 5' and 3' regions ($OR_{TxFlnk}$ = 0.45, p-value$_{TxFlnk}$ = 2.69e-129). Regarding inactive chromatin states, we observed that both CpGs in positive and inverse eQTMs were enriched for quiescent regions ($OR_{Quies}$ = 1.33, p-value$_{Quies}$ = 2.27e-131), while specifically CpGs in positive eQTMs were enriched for repressed and weak repressed Polycomb regions ($OR_{ReprPC}$ = 1.53, p-value$_{ReprPC}$ = 2.17e-114 and $OR_{ReprPCWk}$ = 1.62, p-value$_{ReprPCWk}$ = 1.01e-148, respectively). Finally, enrichment for bivalent regions was dependent on the direction of the effect: while CpGs in positive eQTMs were enriched for flanking bivalent regions ($OR_{BivFlnx}$ = 1.34, p-value$_{BivFlnx}$ = 2.17e-38) and bivalent enhancers ($OR_{EnhBiv}$ = 1.66, p-value$_{EnhBiv}$ = 2.79e-132), CpGs in inverse eQTMs were depleted for bivalent TSSs ($OR_{TssBiv}$ = 0.85, p-value$_{TssBiv}$ = 4.0e-10). Overall, these results suggest that CpGs in eQTMs tend to be in active blood regulatory regions, with CpGs in inverse eQTMs specifically located in promoters (TssAFlnk).

Given that methylation levels characterize regulatory elements, we also evaluated whether methylation levels were related to the presence of eQTMs. We found that CpGs in eQTMs were enriched for CpGs classified as medium methylation ($OR_{Medium\text{-}Met}$ = 2.62, p-value$_{Medium\text{-}Met}$ < 2e-16) and depleted for CpGs classified as low methylation ($OR_{Low\text{-}Met}$ = 0.61, p-value$_{Low\text{-}Met}$ < 2e-16) (Figure 3C; Figure S7C for proportions). We observed the same pattern for inverse and positive-CpGs. To evaluate the influence of methylation levels in the previous

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

enrichment for regulatory elements, we compared the median methylation levels between CpGs in eQTMs versus CpGs not being part of eQTMs, stratified by regulatory element. In all regulatory elements, CpGs in eQTMs tended to exhibit more intermediate methylation levels compared to CpGs not in eQTMs (Figure S8-S9). This was especially evident for CpG islands, active TSSs (TssA), and transcription at 5' and 3' regions (TxFlnk). All these three regulatory elements were depleted among CpGs in eQTMs (Figure 3).

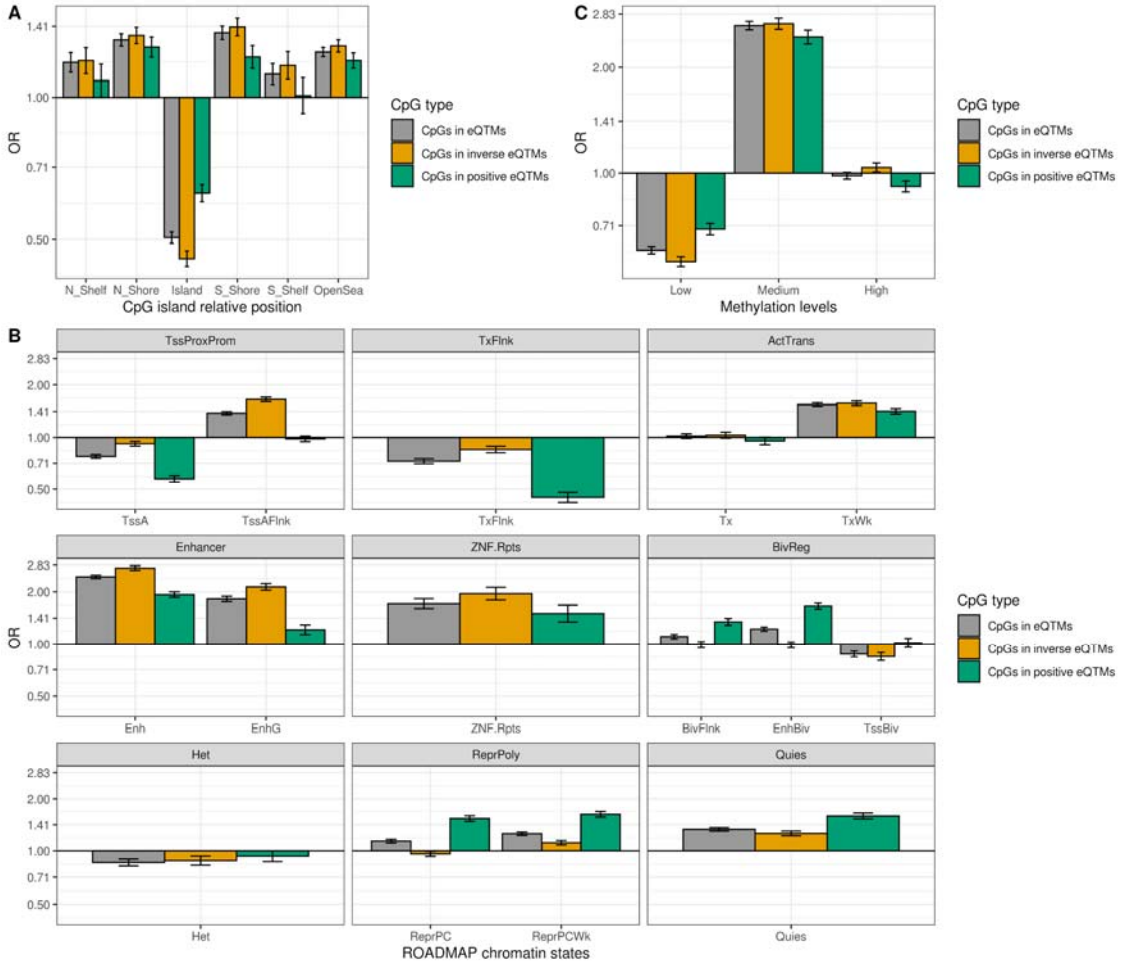



**Figure 3. Enrichment of CpGs in child blood autosomal cis-eQTMs for different regulatory elements.** CpGs were classified in all CpGs in eQTMs (grey); CpGs in inverse eQTMs (yellow); and CpGs in positive eQTMs (green). The y-axis represents the odds ratio (OR) of the enrichment. A) Enrichment for CpG island relative positions: CpG island, N- and S-shore, N- and S-shelf, and open sea. B) Enrichment for ROADMAP blood chromatin states: active TSS (TssA), flanking active TSS (TssAFlnk), transcription at 5' and 3' (TxFlnk), transcription region (Tx), weak transcription region (TxWk), enhancer (Enh); genic enhancer (EnhG), zinc finger genes and repeats (ZNF.Rpts), flanking bivalent region (BivFlnx), bivalent enhancer (EnhBiv), bivalent TSS (TssBiv), heterochromatin (Het), repressed Polycomb (ReprPC), weak repressed polycomb (ReprPCWk), and quiescent region (Quies). C) Enrichment for groups of CpGs with different median methylation levels: low (0-0.3), medium (0.3-0.7), and high (0.7-1)[22].

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Genes in eQTMs are involved in immune system functions

To identify which biological functions were regulated by eQTMs, we ran a gene-set enrichment analysis based on the genes annotated to TCs in these eQTMs. 6,675 out of the 11,071 annotated genes in eQTMs were present in gene ontology (GO), leading to 76 enriched GO terms (q-value < 0.001) (Table S1). As expected, due to the tissue analyzed, 59.2% of the GO terms were related to immune responses (N = 45), followed by GO terms associated with cellular processes (N = 19), and metabolic processes (N = 12). Among immune GO terms, 20 of them were part of innate immunity, 18 of adaptive response and 7 were general/other immune pathways.

## CpGs in eQTMs are enriched for CpGs associated with phenotypic traits and/or environmental exposures

We assessed whether CpGs in eQTMs were enriched for CpGs whose methylation levels had been related to phenotypic traits and/or environmental exposures. To this end, we retrieved CpGs from EWAS performed in blood of European ancestry subjects: 143,384 CpGs from the EWAS catalogue [5], and 54,599 CpGs from the EWAS Atlas [6]. Among them, 16,083 and 9,547 CpGs were part of our eQTMs, representing 45.7% and 27.1% of all eQTMs. We found that CpGs in eQTMs were enriched for CpGs in these EWAS databases in comparison to CpGs not in eQTMs ($OR_{EWAS\text{-}catalogue}$ = 1.48; $p.value_{EWAS\text{-}catalogue}$ < 2e-16; $OR_{EWAS\text{-}Atlas}$ = 2.53; $p\text{-}value_{EWAS\text{-}Atlas}$ < 2-16) (Figure S10). Enrichment was more pronounced in CpGs in inverse eQTMs than in CpGs in positive eQTMs. Of note, CpGs present in the EWAS catalogue and the EWAS Atlas tended to exhibit medium methylation levels compared to CpGs not listed in the datasets (Figure S11). Among them, CpGs in eQTMs did not have a different distribution of methylation levels compared to CpGs not in eQTMs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Annotating CpGs to the closest gene only partially captures eQTMs

A standard approach to interpret EWAS findings is to assume that CpGs regulate the expression of proximal genes. These proximal genes are usually identified through the Illumina 450K annotation [26], which annotates a CpG to a gene when the CpG maps into the gene body, untranslated regions, or promoter region defined as <1,500 bp upstream the TSS. We evaluated to which extent the Illumina 450K annotation captures the eQTMs identified in our catalogue.

To do so, we subsetted 351,909 CpG-TC pairs involving 27,610 unique CpGs and 16,957 unique genes that were present in the Illumina 450K and the Affymetrix HTA 2.0, and thus which could be compared (Table S2). First, we analyzed whether these CpG-TC pairs were more likely to be eQTMs than CpG-TC pairs not annotated to the same gene. As expected, since eQTMs tend to be close to the TSS, 11,675 of the 351,909 CpG-TC pairs were eQTMs (OR = 8.69, p-value < 2e-16). These 11,675 eQTMs represented 25.8% of eQTMs where CpG and TC were annotated to a comparable gene and 18% of the total number of eQTMs. Second, we did a similar comparison but at the CpG level. For 53.6% of the 27,610 CpGs annotated to genes assessed in both platforms (38.4% of the CpGs in all eQTMs), the gene annotated by Illumina coincided with one of the genes (TCs) associated with the CpG through the eQTM analysis. These results suggest that eQTM prediction based on the closest gene misses around one-half of the comparable methylation-expression associations. In other words, in one half of the eQTMs the CpG is located at >1,500 bp from the regulated gene, and thus not captured by the Illumina annotation. Other studies have also found that a substantial part of CpG-gene associations do not involve the nearest gene [27].

Among the 11,675 eQTMs where CpG-TC pairs were annotated to the same gene, we analyzed whether the relative position of the CpG in the genic region was related to the

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expression of the eQTM gene. We used the 351,909 CpG-TC pairs annotated to the same gene as the background. CpGs in these eQTMs were enriched for 5'UTRs ($OR_{5'UTR}$ = 1.10, p-value$_{5'UTR}$ = 9.1e-5) and gene body positions ($OR_{GeneBody}$ = 1.33 p-value$_{GeneBody}$ = 1.6e-55), while depleted for proximal promoter ($OR_{TSS200}$ = 0.68, p-value$_{TSS200}$ = 2.5e-42), first exon ($OR_{1stExon}$ = 0.66, p-value$_{1stExon}$ = 6e-31) and 3'UTRs ($OR_{3'UTR}$ = 0.66, p-value$_{3'UTR}$ = 7.7e-17) (Figure S12). Interestingly, when splitting by the direction of the effect, we observed that CpGs in inverse and positive eQTMs had different behaviors: CpGs in inverse eQTMs were specifically enriched for distal promoter (TSS1500) and 5'UTR, while CpGs in positive eQTMs were enriched for gene body (Figure S12). These results confirm previous knowledge of the effect of CpG gene relative position on the positive/inverse correlations between methylation and gene expression [11].

## Genetic contribution to child blood autosomal cis-eQTM regulation

We hypothesized that genetic variation might regulate DNA methylation and gene expression in a fraction of child blood autosomal cis-eQTMs. To test this, we used two measures of genetic influence: (1) blood methylation heritability for each CpG calculated from twin designs (total additive heritability) and genetic relationship matrices (SNP heritability) as reported by Van Dongen and colleagues [28], and (2) meQTLs (methylation quantitative trait loci, SNPs associated with DNA methylation levels) identified in the ARIES dataset [29].

First, we found that CpGs in eQTMs had higher total additive and SNP heritabilities than CpGs not in eQTMs (with a median difference of 0.23 and 0.10, respectively, and a p-value < 2e-16 for both) (Figure 4A and 4B). Moreover, heritabilities were higher in CpGs with a higher number of associated TCs (for each associated TC, total additive and SNP heritabilities increased 0.025 and 0.026 points, respectively, with a p-value < 2e-16 for both).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

These results suggest that CpGs that regulate the expression of several genes, master regulators, are more prone to be themselves regulated by genetic variation.

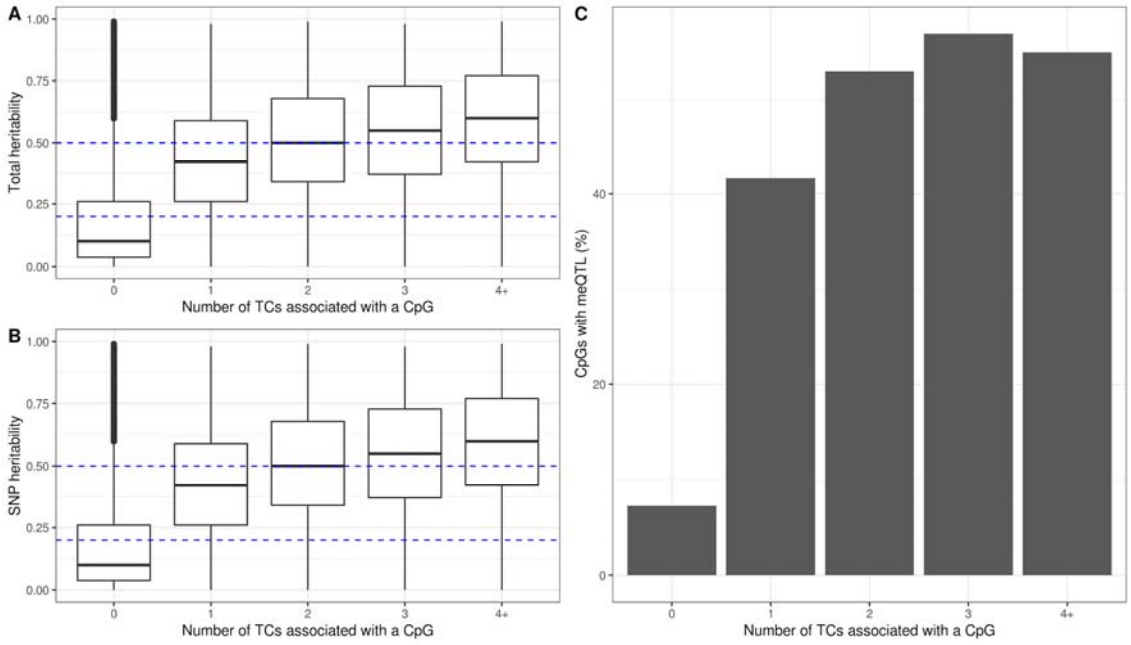



**Figure 4. Genetic contribution to child blood autosomal cis-eQTMs.** CpGs were grouped by the number of TCs they were associated with, where 0 means that a CpG was not associated with any TC (non-eQTMs). A) Total additive heritability as inferred by Van Dongen and colleagues [28], by each group of CpGs associated with a given number of TCs. B) SNP heritability as inferred by Van Dongen and colleagues [28], by each group of CpGs associated with a given number of TCs. C) Proportion of CpGs having a meQTL (methylation quantitative trait locus), by each group of CpGs associated with a given number of TCs.

Second, we studied whether CpGs in eQTMs were enriched for meQTLs, either cis or trans. We restricted our analysis to 2,820,145 meQTLs identified in blood samples of children aged 7 years in the ARIES dataset and replicated in HELIX (see Material and Methods). These 2,820,145 SNP-CpG pairs comprised 36,671 CpGs, of which 10,570 were in eQTMs (30.0% of the CpGs in eQTMs) ($OR_{meQTL}$ = 5.35, p-value$_{meQTL}$ < 2e-16). CpGs regulated by SNPs (meQTLs) accounted for 32.4% of all eQTMs (20,672 eQTMs). While CpGs not in eQTMs were associated with a median of 32 SNPs (IQR = 10; 78), CpGs in eQTMs were associated with a median of 69 (IQR = 24; 156). Moreover, CpGs associated with several TCs (multi-CpGs) were more likely to have a meQTL ($OR_{meQTL}$ = 6.54, p-value$_{meQTL}$ < 2e-16) than CpGs associated with one meQTL (mono-CpGs) ($OR_{meQTL}$ = 4.67, p-value$_{meQTL}$ < 2e-16, Figure

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4C). Overall, we found that a substantial fraction of CpGs associated with gene expression tended to be under genetic control.

Given these findings, we then determined whether SNPs in meQTLs were also eQTLs for the genes of the eQTMs. After multiple-testing correction, we identified 1,305,417 SNP-CpG-TC trios with consistent direction of effect. These trios comprised 15,261 eQTMs (23.9% of total eQTMs), 8,159 unique CpGs (23.2% of CpGs in eQTMs), and 4,247 unique TCs (38.4% of TCs in eQTMs) of which 3,275 were coding (41.6% of coding TCs in eQTMs). In these trios, TCs were associated with a median of 2 CpGs (IQR = 1; 4) and 62 SNPs (IQR = 19; 145), while CpGs were associated with a median of 1 TC (IQR = 1; 2) and 51 SNPs (IQR = 17; 122). Both CpGs and TCs were associated with a considerable number of SNPs, likely due to the linkage disequilibrium. One example of such a SNP-CpG-TC trio is formed by rs11585123-cg15580684-TC01000080.hg.1 (*AJAP1*), in chromosome 10 (Figure S13).

Next, we run a gene-set enrichment analysis with the 2,738 genes involved in these trios. We identified 26 significant GO terms (q-value < 0.001), of which 12 were related to metabolic processes, 8 with immunity (6 innate, 1 adaptive immunity, and 1 general/other), and 6 with cellular processes (Table S3). In contrast to the gene-set analysis performed with genes of all eQTMs, genes of eQTMs under genetic control seem to be deflected to metabolic processes (46.2% vs. 15.8%) rather than to immunity pathways (30.8% vs. 57.9%) (Table S4).

## Effect of blood cellular composition on child autosomal cis-eQTMs

Blood is composed of different cell types that may exhibit different DNA methylation and gene expression patterns. To identify potential cell type-specific eQTMs, we repeated the analyses additionally adjusting for the proportions of the six main blood cell types estimated from the methylation data. We hypothesized that methylation, expression, and cell type

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

proportions will be correlated in cell type-specific eQTMs, but not in eQTMs shared among cell types, and therefore eQTMs significant in the main model but not in the adjusted model would be potential cell type-specific eQTMs (Figure S14).

After adjusting for blood cellular composition, the number of eQTMs decreased from 63,831 to 39,749 (37.7% reduction) (Table 3). Most of these 39,749 eQTMs were already detected in the main model unadjusted for cell type proportions, with only 17.9% being novel eQTMs. Moreover, in the model adjusted for cellular composition, the number of unique CpGs in eQTMs was also reduced substantially (37.6%), while this reduction was less dramatic for TCs (19.7%). Thus, while CpGs were associated with a similar number of TCs in both models, TCs were associated with a lower number of CpGs after adjustment for cell type composition, indicating a loss in the transcriptional complexity (Figure S15).

**Table 3. Comparison of the number of child blood autosomal cis-eQTMs between the main model and the model additionally adjusted for cellular composition.**

| | Main model (unadjusted for cellular composition) | Additional model (adjusted for cellular composition) | Shared between models | Specific to main model (unadjusted for cellular composition) | Specific to additional model (adjusted for cellular composition) |
|---|---|---|---|---|---|
| Autosomal cis-eQTMs | 63,831 | 39,749 | 32,625 | 31,206 (48.9%) | 7,124 (17.9%) |
| TCs | 11,071 | 8,886 | 7,880 | 3,191 (28.8%) | 1,006 (11.3%) |
| Coding TCs | 7,874 | 6,288 | 5,698 | 2,176 (27.6%) | 590 (9.4%) |

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

| | | | | | |
|---|---|---|---|---|---|
| CpGs | 35,228 | 21,966 | 18,529 | 16,699 (47.4%) | 3,437 (15.6%) |

Percentages are referred to the total number of eQTMs, TCs or CpGs for a given model.

We compared the effect estimates of eQTMs between the two models. For eQTMs significant in both models (model-shared eQTMs, N = 32,625), Pearson's correlation of the effect sizes was very high (r = 0.97, p-value < 2e-16) (Figure 5A). Pearson's correlation for eQTMs significant only after adjusting for cellular composition (adjusted cell-specific eQTMs, N = 7,124) was lower (r = 0.8, p-value < 2e-16), but the estimates were still comparable, and they were marginally significant in the other model (Figure 5B). In contrast, Pearson's correlation for eQTMs uniquely found in the main model (unadjusted cell-specific eQTMs, N = 31,206) was much lower (r = 0.54, p-value < 2e-16) with many eQTMs with effect sizes close to zero in the adjusted model (Figure 5B).

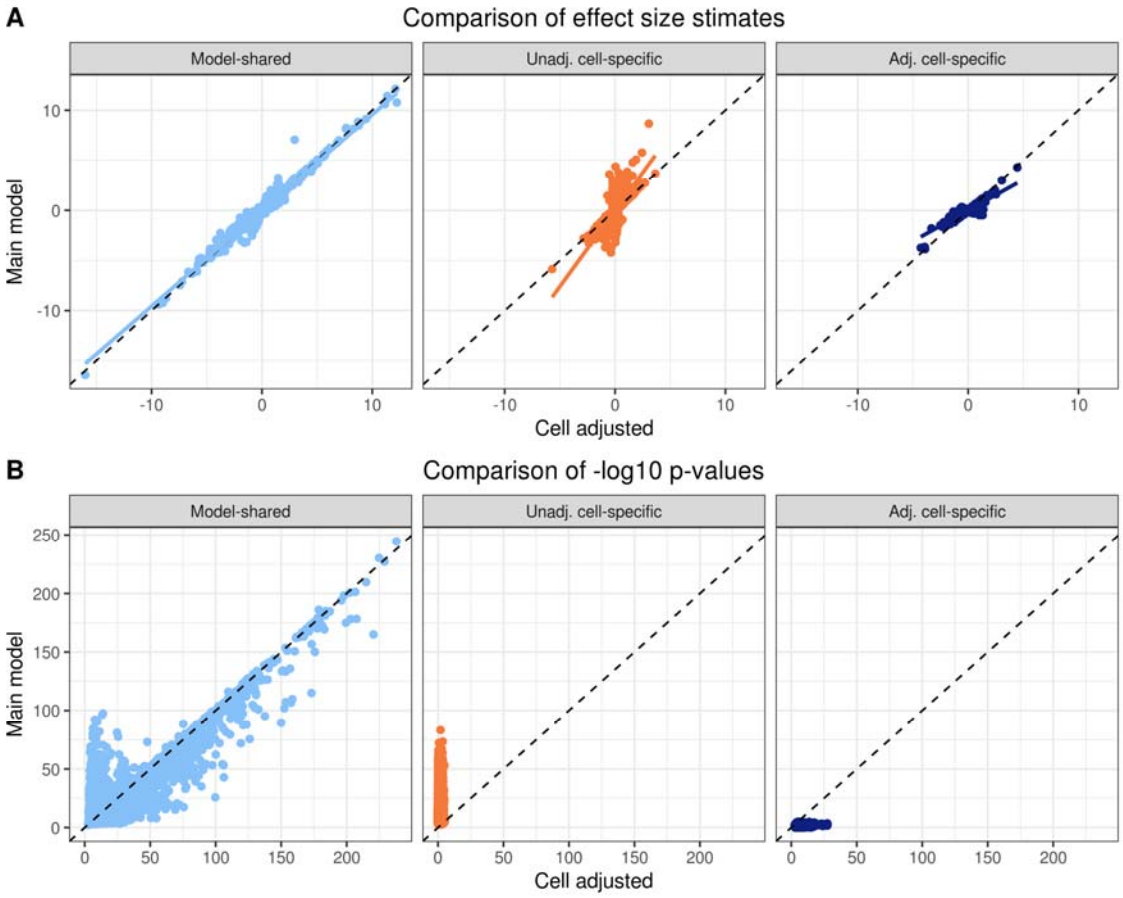



**Figure 5. Effect of blood cellular composition on child autosomal cis-eQTMs: comparison**

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

**between main model and model additionally adjusted for cellular composition.** eQTMs were classified in: model-shared eQTMs (eQTMs identified in both models, in light blue); unadjusted cell-specific eQTMs (eQTMs only identified in the main model unadjusted for cellular composition, in orange); and adjusted cell-specific eQTMs (eQTMs only identified in the model adjusted for cellular composition, in dark blue). A) Comparison of effect estimates. Dots represent the effect size of eQTMs, while dashed black line represents a theoretical regression line if estimates from both models were identical. B) Comparison of $-\log_{10}$ p-values. Dots represent the $-\log_{10}$ p-values of eQTMs, while dashed black line represents a theoretical regression line if $-\log_{10}$ p-values from both models were identical.

Subsequently, we compared the characteristics of model-shared versus unadjusted cell-specific eQTMs. CpGs in model-shared eQTMs were more proximal to the TC TSSs than CpGs in unadjusted cell-specific eQTMs (median $\text{distance}_{\text{model-shared}}$ = 1.1 kb (IQR: -30; 62), median $\text{distance}_{\text{unadj. cell-specific}}$ = 2.7 kb (IQR: -193; 206), p-value < 2e-16) (Figure S16). As model-shared eQTMs were closer to the TSS, in 24.4% of them (N = 7,948) the Illumina annotated gene matched the Affymetrix gene, while this was reduced to 11.9% in unadjusted cell-specific eQTMs (N = 3,727). We also compared whether CpGs in model-shared and unadjusted cell-specific eQTMs were enriched for different ROADMAP regulatory regions (Figure S17). Compared to CpGs not in eQTMs, both CpGs in model-shared and unadjusted cell-specific eQTMs were enriched for enhancers ($\text{OR}_{\text{Enh (model-shared)}}$ = 2.62, $\text{p-value}_{\text{Enh (model-shared)}}$ < 2e-16; $\text{OR}_{\text{Enh (unadj. cell-specific)}}$ = 2.11, $\text{p-value}_{\text{Enh (unadj. cell-specific)}}$ < 2e-16; $\text{OR}_{\text{EnhG (model-shared)}}$ = 1.63, $\text{p-value}_{\text{EnhG (model-shared)}}$ < 2e-16; $\text{OR}_{\text{EnhG (unadj. cell-specific)}}$ = 2.23, $\text{p-value}_{\text{EnhG (unadj. cell-specific)}}$ < 2e-16). In contrast, CpGs in model-shared eQTMs, which were closer to the TSS, were enriched for active states around the promoter ($\text{OR}_{\text{TssA}}$ = 1.14, $\text{p-value}_{\text{TssA}}$ < 2e-16; $\text{OR}_{\text{TssAFlnk}}$ = 1.93, $\text{p-value}_{\text{TssAFlnk}}$ < 2e-16), zinc finger genes and repeats ($\text{OR}_{\text{ZNF.Rpts}}$ = 2.44, $\text{p-value}_{\text{ZNF.Rpts}}$ < 2e-16) and three bivalent regions; while CpGs in unadjusted cell-specific eQTMs were depleted for these same regions ($\text{OR}_{\text{TssA}}$ = 0.44, $\text{p-value}_{\text{TssA}}$ < 2e-16; $\text{OR}_{\text{TssAFlnk}}$ = 0.95, $\text{p-value}_{\text{TssAFlnk}}$ = 1.31e-3; $\text{OR}_{\text{ZNF.Rpts}}$ = 0.94, $\text{p-value}_{\text{ZNF.Rpts}}$ = 0.32). Finally, unadjusted cell-specific eQTMs were uniquely enriched for transcription regions ($\text{OR}_{\text{Tx}}$ = 1.20, $\text{p-value}_{\text{Tx}}$ < 2e-16) and repressed regions ($\text{OR}_{\text{RprPCWk}}$ = 1.60, $\text{p-value}_{\text{RprPCWk}}$ < 2e-16; $\text{OR}_{\text{Quies}}$ = 1.62, $\text{p-value}_{\text{Quies}}$ < 2e-16). The proportions of CpGs in model-shared and unadjusted cell-specific eQTMs annotated to the different ROADMAP chromatin states can be found in Figure S18.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, we checked our hypothesis that CpGs and TCs in eQTMs uniquely identified in the main model (unadjusted cell-specific eQTMs) are blood cell type-specific by contrasting them with data from sorted blood cell types. For this, we retrieved DNA methylation levels from six sorted blood cell types by Reinius and colleagues [30] and gene expression levels from twelve sorted blood cell types from the Blueprint project [31]. We used the $\log_{10}$ of the F-statistic of a linear regression (see Material and Methods) as a measure of cell type specificity (higher F-statistic, higher specificity), as described elsewhere [32]. CpGs in unadjusted cell-specific eQTMs had higher F-statistics than CpGs in model-shared eQTMs (mean change in $\log_{10}$ F-statistic = 0.42, p-value < 2e-16) (Figure S19A). Similarly, genes in unadjusted cell-specific eQTMs (N = 2,278) had higher F-statistics than genes in model-shared eQTMs (N = 5,648) (mean change in $\log_{10}$ F-statistic = 0.10, p-value = 4.57e-11) (Figure S19B). This suggests that unadjusted cell-specific eQTMs, eQTMs uniquely found in the unadjusted model, likely represent blood cell type-specific CpG-TC associations.

## Influence of age on child blood autosomal cis-eQTMs

To understand the association between methylation and gene expression throughout life, we evaluated whether child blood autosomal cis-eQTMs were enriched for CpGs with variable blood methylation levels during childhood/adolescence. To this end, we used two databases: 14,150 CpGs from the MeDALL project whose methylation levels varied between 0 and 8 years (9,647 CpGs with increased and 4,503 CpGs with decreased methylation) [33]; and 244,283 CpGs from the Epidelta project with variable methylation levels between 0 and 17 years (168,314 with increased and 75,969 with decreased methylation) [34]. Of notice, 90% of the CpGs identified in MeDALL were also reported in the Epidelta project. CpGs in eQTMs were enriched for age variable CpGs in both databases ($OR_{MeDALL}$ = 3.69, p-$value_{MeDALL}$ < 2e-16; and $OR_{Epidelta}$ = 2.08, p-$value_{Epidelta}$ < 2e-16), both for CpGs with increased methylation over age ($OR_{MeDALL}$ = 3.30, p-$value_{MeDALL}$ < 2e-16; $OR_{Epidelta}$ = 1.48, p-$value_{Epidelta}$ < 2e-16), as well as decreased ($OR_{MeDALL}$ = 3.72, p-$value_{MeDALL}$ < 2e-16; $OR_{Epidelta}$ = 1.61, p-$value_{Epidelta}$ <

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2e-16) (Figure 6A). CpGs positively associated with gene expression tended to encompass CpGs showing increased methylation over age in both databases, and vice-versa. This suggests that the change in DNA methylation levels in age variable CpGs, either increase or decrease, is related to activation of expression of genes, rather than to repression. For these CpGs, we did not observe different distributions of median methylation levels between CpGs in eQTMs and CpGs not in eQTMs (Figure S20).

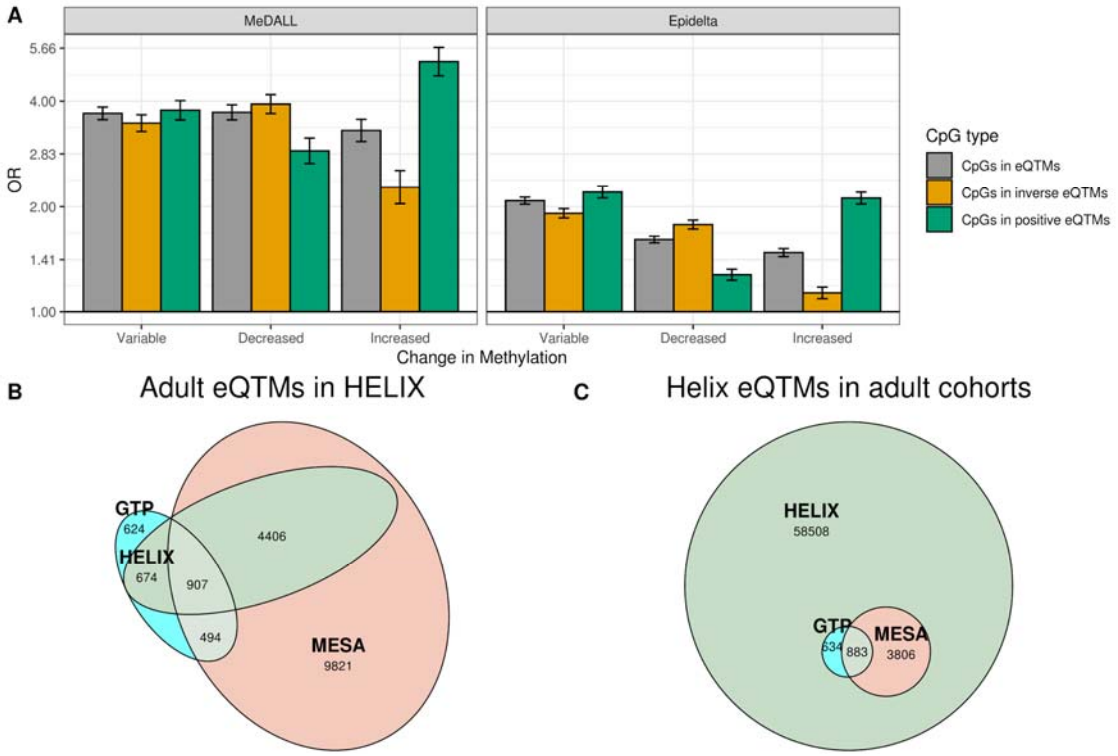



**Figure 6. Influence of age on child blood autosomal cis-eQTMs.** A) Enrichment of CpGs in child blood autosomal cis-eQTMs for CpGs with variable methylation levels. CpGs in eQTMs were classified in all CpGs in eQTMs (grey); CpGs in inverse eQTMs (yellow); and CpGs in positive eQTMs (green). Age variable CpGs were retrieved from the MeDALL project (from birth to childhood [33]) and the Epidelta project (from birth to adolescence [34]), and they were classified in: variable (CpGs whose methylation change over time); decreased (CpGs whose methylation decreases over time); and increased (CpGs whose methylation increases over time). The y-axis represents the odds ratio (OR) of the enrichment. B) Overlap between autosomal cis-eQTMs identified in adults (GTP: whole blood; MESA: monocytes)[11] with eQTMs identified in children (HELIX). All CpG-gene pairs reported at p-value < 1e-5 in GTP or MESA that could be compared with pairs in HELIX are shown. C) Overlap between blood autosomal cis-eQTMs identified in children (HELIX) with eQTMs identified in adults (GTP: whole blood; MESA: monocytes)[11]. All CpG-gene pairs in HELIX that could be compared with pairs in GTP or MESA are shown. Note: The comparison has been split into two plots because one TC in HELIX can be mapped to different expression probes in GTP and MESA and vice-versa. Only comparable CpG-TC pairs are shown (see *Material and Methods*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Subsequently, we evaluated whether child blood autosomal cis-eQTMs were consistent along the life-course. For this, we used data from the adult blood eQTM catalogue of Kennedy and colleagues [11], which, similar to ours and in contrast to the catalogue of Bonder and colleagues [8], includes eQTMs with underlying genetic variation. The catalogue of Kennedy and colleagues contains the summary statistics of all autosomal cis and trans CpG-gene pairs at p-value <1e-05, although only CpG-gene associations at p-value <1e-11 are considered significant eQTMs in their manuscript. For the comparison with our findings we mapped TCs to gene symbols and split the catalogue of Kennedy and colleagues in two, each one representing the findings in one of the adult study populations: (1) GTP (whole blood, 333 samples, 67,606 CpG-gene pairs with p-value < 1e-5 and 2,466 with p-value < 1e-11); and (2) MESA (monocytes, 1,202 samples, 327,049 CpG-gene pairs with p-value < 1e-5 and 34,518 with p-value < 1e-11).

The GTP catalogue contains 2,699 CpG-gene pairs that were also tested in HELIX, and thus comparable. Of those, 1,581 were eQTMs in HELIX (58.6% of all comparable pairs) (Figure 6B). Their effect sizes were correlated (r = 0.37, p-value = 1.59e-52) and 95.6% of them had consistent direction of the effect. When restricting the comparison to eQTMs (CpG-gene pairs with p-value < 1e-11), we obtained similar results: their effect sizes were correlated (r = 0.42, p-value = 1.15e-24) and 94.1% of them showed consistent direction of the effect. On the other hand, 1,118 CpG-gene pairs present in GTP catalogue were not eQTMs in HELIX (41.4%). Their effect sizes showed weaker correlation (r = 0.14, p-value = 2.28e-6) and only 70.4% of them had consistent direction of effect.

The MESA catalogue has 15,628 CpG-gene pairs also analyzed in HELIX. Of those, 5,313 were reported as eQTMs in HELIX (34.0% of all comparable pairs) (Figure 6B), which had correlated effect size (r = 0.31, p-value = 9.47e-116) and for 93.0% of them the direction of the effect was consistent. The weaker correlation of the effect sizes can be explained by cellular composition (monocytes in MESA vs. whole blood in HELIX). Again, these values were similar when restricting the comparison to eQTMs in MESA (r = 0.28, p-value = 5.06e-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

24, 93.6% CpG-gene pairs with consistent direction of the effect). Finally, 10,315 CpG-gene pairs present in MESA catalogue were not eQTMs in HELIX (66.0% of all comparable pairs). As in GTP catalogue, their effect sizes showed lower correlation (r = 0.18, p-value = 4.42e-78) and a lower proportion of CpG-gene pairs had a consistent direction of the effect (72.1%).

Only 5,323 (8.3%) of the eQTMs identified in HELIX children were reported in GTP or MESA catalogues (Figure 6C). We explored whether blood eQTMs identified in adults and children (age-shared eQTMs) had different characteristics compared to blood eQTMs only found in HELIX children (child-specific eQTMs). Age-shared eQTMs involved 4,239 unique CpGs and 1,681 unique TCs, while child-specific eQTMs involved 32,996 unique CpGs and 10,716 unique TCs. 2,007 and 1,326 of those CpGs and TCs, respectively, were part of both types of eQTMs. CpGs in age-shared eQTMs were closer to the TSS compared to child-specific eQTMs (median $distance_{age\text{-}shared}$ = 1.17 kb (IQR = -1.85; 33.9), median $distance_{child\text{-}specific}$ = 1.38 kb (IQR = -99.2; 129.4), p-value < 2e-16)(Figure S21). Also, we observed that CpGs in child-specific eQTMs had higher blood cell type specificity compared to age-shared eQTMs (p-value < 2e-16, Figure S22A), but no major differences were observed at the gene expression level (p-value = 0.039, Figure S22B). Both types of eQTMs were enriched for meQTLs, but enrichment was stronger for CpGs in age-shared eQTMs ($OR_{meQTLs\ (age\text{-}shared)}$ = 19.22, $p\text{-}value_{meQTLs\ (age\text{-}shared)}$ < 2e-16 and $OR_{meQTLs\ (child\text{-}specific)}$ = 4.99, $p\text{-}value_{meQTLs\ (child\text{-}specific)}$ < 2e-16). The enrichment for ROADMAP blood chromatin states was quite similar between the two groups (Figure S23), except for CpGs in age-shared eQTMs which were specifically enriched for active promoter states ($OR_{TssA}$ = 1.34, $p\text{-}value_{TssA}$ < 2e-16) and flanking transcription ($OR_{TssAFlnk}$ = 2.73, $p\text{-}value_{TssAFlnk}$ < 2e-16); while child-specific eQTMs were specifically enriched for repressed ($OR_{ReprPC}$ = 1.29, $p\text{-}value_{ReprPC}$ < 2e-16) and quiescent regions ($OR_{Quies}$ = 1.33, $p\text{-}value_{Quies}$ < 2e-16). Proportions of CpGs in each ROADMAP state can be found in Figure S24. Finally, both types of CpGs were enriched in CpGs variable from birth to childhood/adolescence.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Overall, we found that CpGs in eQTMs were enriched for CpGs whose methylation levels changed from birth to adolescence. Moreover, the overlap between child and adult eQTMs was small: only around 8% of HELIX eQTMs were also described in adults. Child and adult (age-shared) eQTMs tended to be proximal to the TSS, and thus enriched for promoter chromatin states, under the control of genetic variation and common to different blood cells. In contrast, child-specific eQTMs were located at longer distances from the TSS, enriched for repressed regions, and with higher cell type specificity.

# Discussion

In this work, we present the first blood autosomal cis-eQTM catalogue in children. We identified 63,831 eQTMs, representing 35,228 unique CpGs and 11,071 unique TCs. A substantial fraction of these eQTMs was influenced by genetic variation and cellular composition, and the overlap with eQTMs reported in adults was small, indicating that genetics, cellular composition, and age are main factors to be considered in EWAS studies.

The characteristics of the child blood autosomal cis-eQTMs were highly consistent with patterns previously described in other studies. Most of the eQTMs tended to be proximal to the TSS of the gene they were associated with [11,18], but the magnitude of the effect was independent of the distance between the CpG and the TC TSS. Although higher methylation is sometimes assumed to lead to lower expression, we found that around 40% of eQTMs were positively associated with gene expression, a percentage in line with previous results in whole blood (31% [9] and 30% [15]), monocytes (47%)[15], T-cells (31%), lymphoblastoid cell lines (43%) and fibroblasts (49%) from umbilical cord blood [14,15]. CpGs in inverse and positive eQTMs tended to be localized in blood enhancers and other active regulatory regions and not in CpG islands, a pattern also previously reported [9,15]. Despite these common locations, CpGs in inverse eQTMs were specifically found around active TSSs, including the distal promoter and the 5'UTR, while positive-CpGs were localized in gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

body regions. These results highlight the importance of the genomic context to infer the direction of the association of DNA methylation and gene expression. Further studies would be needed to investigate the combined effect of several CpGs located in different genomic regions on the expression of each gene. We note, however, that the causal relationship between DNA methylation and gene expression cannot be ruled out from our study. There is some evidence suggesting that DNA methylation could be a consequence of gene expression, as opposed to the often assumed regulation of gene expression by DNA methylation [14,35].

It has been previously reported that a substantial part of eQTMs is influenced by genetic variation. In HELIX, CpGs in eQTMs showed higher heritabilities, especially if they were linked to several TCs, and 32.4% of the eQTMs involved CpGs for which at least one meQTL was found. Given that genetic variation could be the underlying causal explanation of the association between DNA methylation and gene expression, we searched for SNPs simultaneously associated with DNA methylation and gene expression in our data. We identified 1.3 M SNP-CpG-TC trios with consistent direction of the effect. These SNP-CpG-TC trios are robust as meQTLs were derived the from ARIES database and validated in HELIX, but at the same time this strategy might have missed SNPs less well represented in ARIES or SNPs with stronger effects on gene expression than on DNA methylation. Interestingly, while eQTMs were mostly enriched for immune functions, and to less extent for metabolic and cellular processes, eQTMs under genetic control showed an inverted pattern. This may suggest that the influence of environmental factors is more relevant for immune pathways, while genetic factors might be more determinant in regulating metabolic and cellular processes in blood cells. Given the not negligible effect of genetics in eQTMs, we would advise studying the effect modification of genetic variants on the association between environmental factors and DNA methylation.

DNA methylation and gene expression are cell type-specific [25,30,31,36]. To explore this in our bulk data, we hypothesized that causally related CpG-TC pairs would not be affected by

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

adjustment for cellular composition if their association takes place across cell types, while if it was cell type-specific, it would be attenuated. In consequence, the comparison between unadjusted and adjusted models for cellular composition can provide information about potential cell type-specific eQTMs. Our results seem to support our hypothesis, as eQTMs exclusive to the main unadjusted model, and thus potential cell type-specific, were composed by CpGs and TC with higher cell type specificity (higher F-statistic) in blood cell methylation and expression sorted studies [30,31]. Of note, a high F-statistic can either mean high methylation/expression differences in one particular blood cell type or intermediate methylation/expression differences across several cell types. Besides that, complexity in the transcriptional regulation, assessed as the number of CpGs per TC, was decreased after eliminating the effect of cellular composition. Also, potential cell type-specific eQTMs were located at farther distances from the TSS compared to potential cell shared eQTMs, which were predominantly found in the promoter region. Our findings are consistent with previous literature that describes that promoters act as common regulatory regions across tissues, while enhancers, more distal to the TSS, further tune the expression levels to the needs of each tissue providing additional regulation complexity [37,38]. They are also in line with findings on the regulation of gene expression by genetics, where tissue-shared and tissue-specific eQTLs are enriched for promoters and enhancers, respectively [39]. Nonetheless, additional studies at the single cell level should be performed to further differentiate between shared and cell type-specific eQTMs.

In order to know how eQTMs behave along life-course, we compared blood autosomal cis-eQTMs identified in HELIX children with eQTMs reported by Kennedy and colleagues in whole blood and monocytes from adults [11]. We discarded performing the comparison with the blood eQTMs identified by Bonder and colleagues [8], as their approach has many methodological differences compared to ours, most notably the elimination of the effect of genetic variation on the association of DNA methylation and gene expression. We found that only 8.3% of the child blood eQTMs were also reported in adults. Similarly, a high proportion

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of adult blood eQTMs was neither present in children (41% in GTP and 64% in MESA). This small overlap between adult and child eQTMs has different explanations. Methodological issues such as gene expression platforms with low overlap, statistical methods, and statistical power might have limited the comparison between adults and children in a comprehensive manner. Also, factors other than age, such as cohort-specific environmental exposures or cellular composition might explain the low concordance. Unsurprisingly, HELIX and MESA presented the highest divergence, as HELIX used whole blood and MESA monocytes. Although both HELIX and GTP used whole blood, we cannot discard that differences in eQTMs are a consequence of cell type dynamics over life-course [40]. Despite the effect of these methodological and confounding factors, it is known that DNA methylation and gene expression changes with age [33,34,36], consequently, we expect partial overlap between adult and child eQTMs. The short list of age-shared eQTMs tended to encompass CpGs located in promoters, with lower cell type specificity and highly regulated by genetic variants. Beyond the differences between age groups at the eQTM level, the overall pattern in regulatory elements was similar between adults and children [9,15]. Finally, we also observed that regulatory CpGs in children (CpGs in eQTMs) usually involved CpGs whose methylation varied between birth and childhood/adolescence and that they tended to activate rather than inactivate transcription over this period. They were also enriched for CpGs found to be related to environmental factors and phenotypic traits.

Our catalogue of child blood autosomal cis-eQTMs is meant to improve the biological interpretation of EWAS conducted in children. It has several strengths compared to previous eQTM studies. First, we reported all CpG-TC pairs we tested in our analysis, as opposed to existing blood eQTM catalogues which only reported pairs passing a p-value threshold [8,11]. Reporting all pairs has several advantages: (1) we do not rely on an arbitrary p-value threshold; (2) pairs without association in our study are available for replication and meta-analysis studies, reducing publication bias; (3) researchers can consider pairs not significant in our dataset but with relevant fold change estimates. Second, we reported which eQTMs

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are influenced by genetic variation, and thus researchers can take this into account when exploring the relationship between methylation and expression in their data. In contrast, Kennedy and colleagues did not consider genetic variation as a potential explanatory factor for the eQTMs they described [11], while Bonder and colleagues only reported the association of CpGs with gene expression after removing the effect of genetic variation [8]. Third, the catalogue includes both unadjusted and adjusted models for cellular composition, which might help to identify cell type-specific eQTMs. Overall, and after demonstrating that only half of the CpG-gene relationships are captured through annotation to the closest gene, our eQTM catalogue becomes an essential and powerful tool to help researchers interpret their EWAS studies, with a particular focus on childhood.

The catalogue also has some limitations. First, it only covers a fraction of all CpG-TC pairs, as both the methylation and gene expression arrays have limited resolution. For instance, the methylation array only covers 1-2% of all CpGs in the genome; and the gene expression array although includes >60,000 TCs, coding and non-coding, which is not comparable to untargeted measurements through RNA-seq, at least for not low abundant transcripts. Second, the catalogue does not include sex chromosomes which require more complex analyses to address X-inactivation and sex-specific effects. Third, due to statistical power limitations, only cis effects were tested. Fourth, effect sizes should be considered with caution as the association between DNA methylation and gene expression might be non-linear [41] and effects might be affected by cellular composition. Finally, we have to acknowledge that the catalogue will be useful for biological interpretation of EWAS, if DNA methylation is not a mere mark of cell memory to past exposures without or with time-limited transcriptional consequences [42].

In summary, besides characterizing child blood autosomal cis-eQTMs and how they are affected by genetics, age, and cellular composition, we provide a unique public resource: a catalogue with the association of 13.6 M CpG-gene pairs, with and without adjustment for

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cellular composition, and of 1.3 M SNP-CpG-gene trios (https://helixomics.isglobal.org/). This information will improve the biological interpretation of EWAS findings.

# Methods

## Sample of the study

The Human Early Life Exposome (HELIX) study is a collaborative project across 6 established and on-going longitudinal population-based birth cohort studies in Europe [43]: the Born in Bradford (BiB) study in the UK [44], the Étude des Déterminants pré et postnatals du développement et de la santé de l'Enfant (EDEN) study in France [45], the INfancia y Medio Ambiente (INMA) cohort in Spain [46], the Kaunus cohort (KANC) in Lithuania [47], the Norwegian Mother, Father and Child Cohort Study (MoBa)[48] and the RHEA Mother Child Cohort study in Crete, Greece [49]. All participants in the study signed an ethical consent and the study was approved by the ethical committees of each study area [43].

In the present study, we selected a total of 832 children of European ancestry that had both DNA methylation and gene expression data. Ancestry was determined with cohort-specific self-reported questionnaires.

## Biological samples

DNA was obtained from buffy coat collected in EDTA tubes at mean age 8.1 years old. Briefly, DNA was extracted using the Chemagen kit (Perkin Elmer) in batches of 12 samples. Samples were extracted by cohort. DNA concentration was determined in a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Fisher Scientific) and with Quant-iT™ PicoGreen® dsDNA Assay Kit (Life Technologies).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

RNA was extracted from whole blood samples collected in Tempus tubes (Applied Biosystems) using the MagMAX for Stabilized Blood Tubes RNA Isolation Kit (Thermo Fisher Scientific). RNA extraction was performed in batches of 12-24 samples and by cohort. The quality of RNA was evaluated with a 2100 Bioanalyzer (Agilent) and the concentration with a NanoDrop 1000 UV-Vis Spectrophotometer (Thermo Fisher Scientific). Samples classified as good RNA quality had a RNA Integrity Number (RIN) > 5, a similar RNA integrity pattern at visual inspection, and a concentration >10 ng/ul. Mean values for the RIN, concentration (ng/ul) and Nanodrop 260/230 ratio were: 7.05, 109.07 and 2.15.

## DNA methylation assessment

DNA methylation was assessed with the Infinium HumanMethylation450K BeadChip (Illumina), following manufacturer's protocol at the National Spanish Genotyping Centre (CEGEN), Spain. Briefly, 700 ng of DNA were bisulfite-converted using the EZ 96-DNA methylation kit following the manufacturer's standard protocol, and DNA methylation measured using the Infinium protocol. A HapMap sample was included in each plate. In addition, 24 HELIX inter-plate duplicates were included. Samples were randomized considering cohort, sex, and panel. Samples from the panel study (samples from the same subject collected at different time points) were processed in the same array. Two samples were repeated due to their overall low quality.

DNA methylation data was pre-processed using *minfi* R package [50]. We increased the stringency of the detection p-value threshold to <1e-16, and probes not reaching a 98% call rate were excluded [51]. Two samples were filtered due to overall quality: one had a call rate <98% and the other did not pass quality control parameters of the *MethylAid* R package [52]. Then, data was normalized with the functional normalization method with Noob background subtraction and dye-bias correction [53]. Then, we checked sex consistency using the *shinyMethyl* R package [54], genetic consistency of technical duplicates, panel samples, and other samples making use of the genotype probes included in the Infinium

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

HumanMethylation450K BeadChip and the genome-wide genotyping data when available. In total four samples were excluded, two with discordant sex and two with discordant genotypes. Batch effect (slide) was corrected using the *ComBat* R package [55]. Duplicated samples, one of the samples from the panel study and HapMap samples were removed as well as control probes, probes in sexual chromosomes, probes designed to detect Single Nucleotide Polymorphisms (SNPs) and probes to measure methylation levels at non-CpG sites, giving a final number of 386,518 probes.

CpG annotation was conducted with the *IlluminaHumanMethylation450kanno.ilmn-12.hg19* R package [26]. Briefly, this package annotates CpGs to promoter (up to 1500 bp from TSS), 5'UTR, first exon, gene body, and 3'UTR. CpGs farther than 1,500 bp from the TSS were not annotated to any gene and the promoter region was divided in proximal promoter (200 bp upstream the TSS (TSS200)) and distant promoter (from 200 to 1,500 bp upstream the TSS (TSS1500)). Relative position to CpG islands (island, shelve, shore and open sea) was also provided by the same R package.

Annotation of CpGs to 15 chromatin states was retrieved from the Roadmap Epigenomics Project web portal (https://egg2.wustl.edu/roadmap/web_portal/). Each CpG in the array was annotated to one or several chromatin states by taking a state as present in that locus if it was described in at least 1 of the 27 blood-related cell types.

## Gene expression assessment

Gene expression, including coding and non-coding transcripts, was assessed with the Human Transcriptome Array 2.0 ST arrays (HTA 2.0) (Affymetrix) at the University of Santiago de Compostela (USC), Spain. Amplified and biotinylated sense-strand DNA targets were generated from total RNA. Affymetrix HTA 2.0 arrays were hybridized according to Affymetrix recommendations using the Manual Target preparation for GeneChip Whole Transcript (WT) expression arrays and the labeling and hybridization kits. In each round,

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

several batches of 24-48 samples were processed. Samples were randomized within each batch considering sex and cohort. Samples from the same subject (panel study) were processed in the same batch. Two different types of control RNA samples (HeLa or FirstChoice® Human Brain Reference RNA) were included in each batch, but they were hybridized only in the first batches. Raw data were extracted with the AGCC software (Affymetrix) and stored into CEL files. Ten samples failed during the laboratory process (7 did not have enough cRNA or ss-cDNA, 2 had low fluorescence, and 1 presented an artifact in the CEL file).

Data was normalized with the GCCN (SST-RMA) algorithm at the gene level. Annotation of transcript clusters (TCs) was done with the ExpressionConsole software using the HTA-2.0 Transcript Cluster Annotations Release na36 annotation file from Affymetrix. A TC is defined as a group of one or more probes covering a region of the genome reflecting all the exonic transcription evidence known for the region and corresponding to a known or putative gene. After normalization, several quality control checks were performed and four samples with discordant sex and two with low call rates were excluded [56]. One of the samples from the panel study was also eliminated for this analysis. Control probes and probes in sexual chromosomes or probes without chromosome information were excluded. Probes with a DABG (Detected Above Background) p-value <0.05 were considered to have an expression level different from the background, and they were defined as detected. Probes with a call rate <1% were excluded from the analysis. The final dataset consisted of 58,254 TCs.

Gene expression values were $\log_2$ transformed and batch effect controlled by residualizing the effect of surrogate variables calculated with the sva method [57] while protecting for main variables in the study (cohort, age, sex, and blood cellular composition).

## Blood cellular composition

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Main blood cell type proportions (CD4+ and CD8+ T-cells, natural killer cells, monocytes, eosinophils, neutrophils, and B-cells) were estimated using the Houseman algorithm [58] and the Reinius reference panel [30] from raw methylation data.

## Genome-wide genotyping

Genome-wide genotyping was performed using the Infinium Global Screening Array (GSA) MD version 1 (Illumina), which contains 692,367 variants, at the Human Genomics Facility (HuGe-F), Erasmus MC, The Netherlands. Genotype calling was done using the GenTrain2.0 algorithm based on a custom cluster file implemented in the GenomeStudio software. Annotation was done with the GSAMD-24v1-0_20011747_A4 manifest. Samples were genotyped in two rounds, and 10 duplicates were included which confirmed high inter-round consistency.

Quality control was performed with the PLINK program following standard recommendations [59,60]. We applied the following sample quality controls: sample call rate <97% (N filtered=43), sex concordance (N=8), heterozygosity (N=0), relatedness (N=10, including potential DNA contamination), duplicates (N=19). Then we used the *peddy* python script to predict ancestry from GWAS data [61]. To do so, 6,642 genetic variants, highly polymorphic among populations, and data from the 1000G project were used [62]. We contrasted ancestry predicted from GWAS with ancestry recorded in the questionnaires. Twelve samples were excluded due to discordances between the two variables. Overall, 93 (6.7%) samples, including the duplicates, were filtered out. The variant quality control included the following steps: variant call rate <95% (N filtered=4,046), non-canonical PAR (N=47), minor allele frequency (MAF) <1% (N=178,017), Hardy-Weinberg equilibrium (HWE) p-value <1E-06 (N=913). Note that the QC of sexual chromosomes considered individuals' sex. Some other SNPs were filtered out during the matching between data and reference panel before imputation (N=14,436).

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Imputation of the GWAS data was performed with the Imputation Michigan server [63] using the Haplotype Reference Consortium (HRC) cosmopolitan panel, Version r1.1 2016 [64]. Before imputation, PLINK GWAS data was converted into VCF format and variants were aligned with the reference genome. The phasing of the haplotypes was done with Eagle v2.4 [65] and the imputation with minimac4 [66], both implemented in the code of the Imputation Michigan server. In total, we retrieved 40,405,505 variants after imputation. Then, we applied the following QC criteria to the imputed dataset: imputation accuracy (R2) >0.9, MAF >1%, HWE p-value >1E-06; and genotype probabilities were converted to genotypes using the best guest approach. The final post-imputation quality-controlled dataset consisted of 1,304 samples and 6,143,757 variants (PLINK format, Genome build: GRCh37/hg19, + strand).

## Identification of child blood autosomal cis-eQTMs

To test associations between DNA methylation levels and gene expression levels in cis (cis-eQTMs), we paired each TC to all CpGs closer than 500 Kb from its TSS, either upstream or downstream. In the main analysis, we fitted for each CpG-TC pair a linear regression model between gene expression and methylation levels adjusted for age, sex, and cohort. A second model was run additionally adjusted for blood cellular composition, estimated from DNA methylation data, as described above.

To ensure that CpGs paired to a higher number of TCs do not have higher chances of being part of an eQTM, multiple-testing was controlled at the CpG level, following a procedure previously applied by Bonder and colleagues [8]. To this end, we generated 100 permuted gene expression datasets and ran our previous linear regression models obtaining 100 permuted p-values for each CpG-TC pair. Then, for each CpG, we selected among all CpG-TC pairs the minimum p-value in each permutation and fitted a beta distribution. Next, for each CpG, we took the minimum p-value observed in the real data and used the beta distribution to compute the probability of observing a smaller p-value. This probability was the adjusted p-value at the CpG level. Finally, we considered as significant those CpGs with

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

empirical p-values significant at 5% false discovery rate (FDR), based on Benjamini-Hochberg. Finally, in order to define significant CpG-TC pairs, we selected the CpG with the maximum p-value which was considered as significant and used this adjusted p-value as the significance threshold. Then, we went back to the beta distributions at the CpG level and selected any CpG-TC pair whose p-value was smaller than the significance threshold. We applied the same process for the model adjusted for cellular composition.

## Characterization of the child blood autosomal cis-eQTM catalogue

We used different statistical methods to characterize CpGs and TCs of the eQTM catalogue. A linear regression was run to compare the methylation range vs. methylation levels categories (low, medium, high). Enrichment of CpGs/TCs for regulatory elements were tested using Chi-square tests with CpGs/TCs not in eQTMs as reference, unless otherwise stated. Results with a p-value < 0.05 were considered as significant. Annotation of CpGs to regulatory elements is described in the section “DNA methylation assessment”.

We explored the enrichment of CpGs in eQTMs for phenotypic traits and/or environmental exposures using the EWAS catalogue [5] and the EWAS Atlas [6]. We used version 03-07-2019 of the EWAS catalogue and selected those studies conducted in whole or peripheral blood of European ancestry individuals. We downloaded EWAS Atlas data on 27-11-2019 and selected those studies performed in whole blood or peripheral blood of European ancestry individuals or with unreported ancestry. For both catalogues, we considered all associations from selected studies.

We used results from the MeDALL and the Epidelta projects to test whether CpGs in eQTMs were enriched for CpGs variable from birth to childhood or adolescence, respectively. For MeDALL we downloaded data from supplementary material of the following manuscript [33]. For Epidelta, we downloaded the full catalogue (version 2020-07-17) from their website

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(http://epidelta.mrcieu.ac.uk/). We considered a CpG as variable if its p-value from model 1 (variable M1.change.p) was <1e-7 (Bonferroni threshold as suggested in their manuscript). Variable CpGs were classified as increased methylation if their change estimate (variable M1.change.estimate) was >0, and as decreased methylation otherwise.

We also tested whether genes in eQTMs were enriched for specific GO terms using the *topGO* R package [67]. We analyzed GO terms in the biological processes' ontology using the *weight01* algorithm, which considers GO terms hierarchy for p-values computation. GO terms with q-value < 0.001 were considered as significant.

## Comparison with annotation to close gene

We evaluated whether using annotation of CpGs to the closest gene (Illumina annotation) captured eQTM associations. CpGs were annotated to Gene Symbol using the *IlluminaHumanMethylation450kanno.ilmn-12.hg19* R package [26], while TCs were annotated to Gene Symbol using the HTA-2.0 Transcript Cluster Annotations Release na36 annotation file from Affymetrix. Given that CpGs and TCs could be annotated to several genes, we considered that a CpG-TC pair was annotated to a comparable gene if at least one of the genes annotated to the CpG matched at least one of the genes annotated to the TC. In total, we identified 351,909 comparable CpG-TC pairs. Then, a Chi-square test was run to compute whether these 351,909 comparable CpG-TC pairs were enriched for CpG-TC pairs being eQTMs.

Next, we evaluated whether the relative position of the CpG in the genic region was related to the expression of the eQTM-linked gene. To do so, the comparable 351,909 CpG-TC pairs were expanded to 411,900 entries. Each entry represented a CpG-TC pair annotated to a unique different gene, thus, for instance a CpG-TC pair annotated to two different genes, was included as two entries. In this expanded CpG-TC pair set, Chi-square tests were run to test the enrichment of CpGs in eQTMs for relative gene positions.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Evaluation of the genetic contribution on child blood autosomal cis-eQTMs

We used two approaches to evaluate the influence of genetic effects in child blood autosomal cis-eQTMs. First, we used heritability estimates of CpGs computed by Van Dongen and colleagues [28]. Median total additive and SNP-heritability was compared between CpGs in eQTMs and CpGs not in eQTMs, using a Wilcoxon test. For CpGs in eQTMs, linear regressions between heritability measures (total additive and SNP heritabilities) and the number of TCs associated with each CpG were run.

Second, we tested whether CpGs in eQTMs were more likely regulated by SNPs, thus they were enriched for meQTL. In order to define meQTLs in HELIX, we selected 9.9 M cis and trans SNP-CpG pairs with a p-value < 1e-7 in the ARIES dataset consisting of data from children of 7 years old [29]. In this subset of 9.9 M cis and trans-CpG pairs, we run meQTL analyses using *MatrixEQTL* R package [68], adjusting for cohort, sex, age, blood cellular composition (similar to ARIES) and the first 20 principal components (PCs) calculated from genome-wide genetic data of the GWAS variability. We considered as significant meQTLs the SNP-CpG pairs reaching a p-value < 1e-7 also in HELIX. Enrichment of CpGs in eQTMs for CpGs with meQTLs was computed using a Chi-square test.

Finally, we tested whether meQTLs were also eQTLs for the gene in the eQTM. To this end, we run eQTL analyses with *MatrixEQTL* adjusting for cohort, sex, age, blood cellular composition and the first 20 GWAS PCs in HELIX. We considered as significant eQTLs the SNP-TC pairs with p-value < 1e-7 and with the direction of the effect consistent with the direction of the meQTL and the eQTM.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## Evaluation of the effect of cellular composition on child blood autosomal cis-eQTMs

We compared the results obtained in the main model and in the model additionally adjusted for cellular composition. eQTMs were classified in three groups: eQTMs significant in both models (model-shared eQTMs), eQTMs only significant in the main model (unadjusted cell-specific eQTMs) and eQTMs only significant in the cell adjusted model (adj. cell-specific eQTMs). For each group of eQTMs, fold changes (FC) obtained with each model were compared using a Pearson correlation.

To test if eQTMs were cell type-specific, we used as a proxy of cell type specificity the F-statistic of a linear regression between gene expression/CpG methylation levels in sorted blood cell types (outcome) vs. cell type (predictor), as described before [30]. Higher F-statistics are assumed to be indicative of higher cell type specificity. DNA methylation and gene expression in sorted blood cell types were retrieved from [30] and the Blueprint project [31] (https://blueprint.haem.cam.ac.uk/bloodatlas/), respectively. Once gene/CpG cell type specificity was calculated and $\log_{10}$ transformed ($\log_{10}$ F-statistic), we tested its association with the eQTM category (model-shared or unadjusted cell-specific eQTMs) by fitting linear regressions.

## Comparison with adult blood eQTM catalogues: GTP and MESA

We compared our list of child blood autosomal cis-eQTMs obtained in HELIX with the eQTMs described in blood of two adult cohorts: GTP and MESA [11]. DNA methylation was assessed with the Infinium HumanMethylation450K BeadChip (Illumina) in the 3 cohorts (HELIX, GTP and MESA). In HELIX, gene expression was assessed with the Human

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transcriptome Array 2.0 ST arrays (HTA 2.0) (Affymetrix), and in GTP and MESA with the HumanHT-12 v3.0 and v4.0 Expression BeadChip (Illumina).

In order to compare eQTMs, the comparison was done at the gene symbol level. In HELIX, we selected the gene symbol of the most likely mRNA mapped to the transcript cluster (TC). In GTP and MESA, we used the gene symbol annotation provided by the authors. As a result of this process, different TCs or expression probes were mapped to the same gene symbol. Thus, a CpG-TC pair in HELIX was mapped to multiple CpG-pairs in GTP and MESA and vice-versa. To handle this issue, we split our comparison in two pairs. First, we checked whether CpG-gene pairs in GTP and MESA were eQTMs in HELIX. When a CpG-gene pair in GTP or MESA mapped to multiple CpG-gene pairs in HELIX, we only considered the CpG-gene pairs with the smallest p-value in HELIX. For the common pairs, Pearson correlations between the effect sizes were computed. Second, we compared those eQTMs in HELIX present in GTP or MESA catalogues with those eQTMs absent in the catalogues. Effect sizes were compared using a Wilcoxon test.

# Author contributions

CR-A and MB designed the study. MV is coordinator of the HELIX project. DM, SC, MC, SA, KBG, MV, JW, JL, RG, LCh recruited participants and obtained biological samples. AC, IQ and MB obtained DNA methylation data; MV-U, EM, XE and MB gene expression data; and GE, KBG and MB genome-wide genotypic data. CR-A, CH-F and GE performed the QC of the omics data. CR-A and SM, under the supervision of MB and JRG, performed the statistical and bioinformatics analyses. CR-A and MB wrote the manuscript and all others approved it.

# Competing interests

The authors declare that they have no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# Acknowledgments

The authors acknowledge the contribution of all the HELIX children and their families.

# Funding

The study has received funding from the European Community's Seventh Framework Programme (FP7/2007-206) under grant agreement no 308333 (HELIX project); the H2020-EU.3.1.2. - Preventing Disease Programme under grant agreement no 874583 (ATHLETE project); from the European Union's Horizon 2020 research and innovation programme under grant agreement no 733206 (LIFECYCLE project), and from the European Joint Programming Initiative "A Healthy Diet for a Healthy Life" (JPI HDHL and Instituto de Salud Carlos III) under the grant agreement no AC18/00006 (NutriPROGRAM project). The genotyping was supported by the project PI17/01225, 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 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

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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). We acknowledge support from the Spanish Ministry of Science and Innovation through the “Centro de Excelencia Severo Ochoa 2019-2023” Program (CEX2018-000806-S), and support from the Generalitat de Catalunya through the CERCA Program.

MV-U and CR-A were supported by a FI fellowship from the Catalan Government (FI-DGR 2015 and #016FI_B 00272). MC received funding from Instituto Carlos III (Ministry of Economy and Competitiveness) (CD12/00563 and MS16/00128).

# References

1. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat. Rev. Genet. Nature Publishing Group; 2017. p. 441–51.

2. Cavalli G, Heard E. Advances in epigenetics link genetics to the environment and disease [Internet]. Nature. Nature Publishing Group; 2019 [cited 2020 Oct 28]. p. 489–99. Available from: https://pubmed.ncbi.nlm.nih.gov/31341302/

3. Feinberg AP. The Key Role of Epigenetics in Human Disease Prevention and Mitigation. N Engl J Med [Internet]. New England Journal of Medicine (NEJM/MMS); 2018 [cited 2020 Oct 28];378:1323–34. Available from: https://pubmed.ncbi.nlm.nih.gov/29617578/

4. Hanson MA, Gluckman PD. Early developmental conditioning of later health and disease: physiology or pathophysiology? [Internet]. Physiol. Rev. Physiol Rev; 2014 [cited 2020 Oct 28]. p. 1027–76. Available from: https://pubmed.ncbi.nlm.nih.gov/25287859/

5. MRC-IEU EWAS Catalog [Internet]. [cited 2020 Oct 28]. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

http://www.ewascatalog.org/

6. Li M, Zou D, Li Z, Gao R, Sang J, Zhang Y, et al. EWAS Atlas: A curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res [Internet]. Oxford University Press; 2019 [cited 2020 Oct 28];47:D983–8. Available from: https://pubmed.ncbi.nlm.nih.gov/30364969/

7. Sharp GC, Salas LA, Monnereau C, Allard C, Yousefi P, Everson TM, et al. Maternal BMI at the start of pregnancy and offspring epigenome-wide DNA methylation: Findings from the pregnancy and childhood epigenetics (PACE) consortium. Hum Mol Genet [Internet]. Oxford University Press; 2017 [cited 2020 Oct 28];26:4067–85. Available from: https://pubmed.ncbi.nlm.nih.gov/29016858/

8. Bonder MJ, Luijk R, Zhernakova D V, Moed M, Deelen P, Vermaat M, et al. Disease variants alter transcription factor levels and methylation of their binding sites. Nat Genet [Internet]. 2017 [cited 2017 Nov 2];49:131–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27918535

9. Küpers LK, Monnereau C, Sharp GC, Yousefi P, Salas LA, Ghantous A, et al. Meta-analysis of epigenome-wide association studies in neonates reveals widespread differential DNA methylation associated with birthweight. Nat Commun [Internet]. Nature Publishing Group; 2019 [cited 2020 Oct 28];10. Available from: https://pubmed.ncbi.nlm.nih.gov/31015461/

10. Gondalia R, Baldassari A, Holliday KM, Justice AE, Méndez-Giráldez R, Stewart JD, et al. Methylome-wide association study provides evidence of particulate matter air pollution-associated DNA methylation. Environ Int [Internet]. Elsevier Ltd; 2019 [cited 2020 Oct 28];132. Available from: https://pubmed.ncbi.nlm.nih.gov/31208937/

11. Kennedy EM, Goehring GN, Nichols MH, Robins C, Mehta D, Klengel T, et al. An integrated -omics analysis of the epigenetic landscape of gene expression in human blood

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cells. BMC Genomics [Internet]. BioMed Central Ltd.; 2018 [cited 2020 Oct 28];19. Available from: https://pubmed.ncbi.nlm.nih.gov/29914364/

12. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De la Fuente A, et al. Methylomics of gene expression in human monocytes. Hum Mol Genet [Internet]. Hum Mol Genet; 2013 [cited 2020 Oct 28];22:5065–74. Available from: https://pubmed.ncbi.nlm.nih.gov/23900078/

13. Husquin LT, Rotival M, Fagny M, Quach H, Zidane N, McEwen LM, et al. Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation 06 Biological Sciences 0604 Genetics. Genome Biol [Internet]. BioMed Central Ltd.; 2018 [cited 2020 Oct 28];19. Available from: https://pubmed.ncbi.nlm.nih.gov/30563547/

14. Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, et al. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. Elife [Internet]. 2013 [cited 2018 Oct 1];2. Available from: https://elifesciences.org/articles/00523

15. Gutierrez-Arcelus M, Ongen H, Lappalainen T, Montgomery SB, Buil A, Yurovsky A, et al. Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing. PLoS Genet [Internet]. Public Library of Science; 2015 [cited 2020 Oct 28];11. Available from: https://pubmed.ncbi.nlm.nih.gov/25634236/

16. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol [Internet]. BioMed Central Ltd.; 2014 [cited 2020 Oct 28];15. Available from: https://pubmed.ncbi.nlm.nih.gov/24555846/

17. Bonder MJ a., Kasela S, Kals M, Tamm R, Lokk K, Barragan I, et al. Genetic and epigenetic regulation of gene expression in fetal and adult human livers. BMC Genomics. 2014;

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

18. Leland Taylor D, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 2019 [cited 2020 Oct 28];166:10883–8. Available from: https://pubmed.ncbi.nlm.nih.gov/31076557/

19. Kim S, Forno E, Zhang R, Park HJ, Xu Z, Yan Q, et al. Expression Quantitative Trait Methylation Analysis Reveals Methylomic Associations With Gene Expression in Childhood Asthma. Chest [Internet]. Elsevier BV; 2020 [cited 2020 Oct 29]; Available from: https://pubmed.ncbi.nlm.nih.gov/32569636/

20. Delahaye F, Do C, Kong Y, Ashkar R, Salas M, Tycko B, et al. Genetic variants influence on the placenta regulatory landscape. PLoS Genet [Internet]. Public Library of Science; 2018 [cited 2020 Oct 29];14. Available from: https://pubmed.ncbi.nlm.nih.gov/30452450/

21. Felix JF, Joubert BR, Baccarelli AA, Sharp GC, Almqvist C, Annesi-Maesano I, et al. Cohort profile: Pregnancy and childhood epigenetics (PACE) consortium. Int J Epidemiol [Internet]. Oxford University Press; 2018 [cited 2020 Oct 29];47:22-23u. Available from: https://pubmed.ncbi.nlm.nih.gov/29025028/

22. Huse SM, Gruppuso PA, Boekelheide K, Sanders JA. Patterns of gene expression and DNA methylation in human fetal and adult liver. BMC Genomics [Internet]. BioMed Central Ltd.; 2015 [cited 2020 Oct 29];16:981. Available from: https://pubmed.ncbi.nlm.nih.gov/26589361/

23. Lin X, Teh AL, Chen L, Lim IY, Tan PF, MacIsaac JL, et al. Choice of surrogate tissue influences neonatal EWAS findings. BMC Med [Internet]. BioMed Central Ltd.; 2017 [cited 2020 Nov 2];15. Available from: https://pubmed.ncbi.nlm.nih.gov/29202839/

24. Gamazon ER, Segrè A V., Van De Bunt M, Wen X, Xi HS, Hormozdiari F, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

associated variation. Nat Genet [Internet]. Nature Publishing Group; 2018 [cited 2020 Oct 29];50:956–67. Available from: https://pubmed.ncbi.nlm.nih.gov/29955180/

25. Roadmap Epigenomics Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature [Internet]. NIH Public Access; 2015 [cited 2018 Feb 21];518:317–30. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25693563

26. Hansen K. IlluminaHumanMethylation450kanno.ilmn12.hg19: Annotation for Illumina's 450k methylation arrays. [Internet]. Available from: https://bioconductor.org/packages/release/data/annotation/html/IlluminaHumanMethylation450kanno.ilmn12.hg19.html

27. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, et al. Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun [Internet]. Nature Publishing Group; 2018 [cited 2020 Oct 29];9. Available from: https://pubmed.ncbi.nlm.nih.gov/29500431/

28. van Dongen J, Nivard MG, Willemsen G, Hottenga J-J, Helmer Q, Dolan C V., et al. Genetic and environmental influences interact with age and sex in shaping the human methylome. Nat Commun [Internet]. Nature Publishing Group; 2016;7:11115. Available from: http://www.nature.com/doifinder/10.1038/ncomms11115

29. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, et al. Systematic identification of genetic influences on methylation across the human life course. Genome Biol [Internet]. Genome Biol; 2016 [cited 2020 Oct 29];17:61. Available from: https://pubmed.ncbi.nlm.nih.gov/27036880/

30. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, et al. Differential DNA methylation in purified human blood cells: Implications for cell lineage and studies on disease susceptibility. Ting AH, editor. PLoS One [Internet]. PLoS One; 2012 [cited 2020 Oct

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29];7:e41361. Available from: https://pubmed.ncbi.nlm.nih.gov/22848472/

31. Grassi L, Izuogu OG, Jorge NAN, Seyres D, Bustamante M, Burden F, et al. Cell type specific novel lncRNAs and circRNAs in the BLUEPRINT haematopoietic transcriptomes atlas. Haematologica [Internet]. Ferrata Storti Foundation (Haematologica); 2020 [cited 2020 Oct 29];haematol.2019.238147. Available from: https://pubmed.ncbi.nlm.nih.gov/32703790/

32. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol [Internet]. BioMed Central Ltd.; 2014 [cited 2020 Oct 29];15. Available from: https://pubmed.ncbi.nlm.nih.gov/24495553/

33. Xu C-J, Bonder MJ, Söderhäll C, Bustamante M, Baïz N, Gehring U, et al. The emerging landscape of dynamic DNA methylation in early childhood. BMC Genomics [Internet]. BioMed Central; 2017 [cited 2017 Aug 3];18:25. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28056824

34. Mulder R, Neumann A, Cecil C, Walton E, Houtepen L, Simpkin A, et al. Epigenome-wide change and variation in DNA methylation from birth to late adolescence. bioRxiv [Internet]. Cold Spring Harbor Laboratory; 2020 [cited 2020 Oct 29];2020.06.09.142620. Available from: https://doi.org/10.1101/2020.06.09.142620

35. Jones PA. Functions of DNA methylation: Islands, start sites, gene bodies and beyond [Internet]. Nat. Rev. Genet. Nat Rev Genet; 2012 [cited 2020 Oct 29]. p. 484–92. Available from: https://pubmed.ncbi.nlm.nih.gov/22641018/

36. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, et al. The human transcriptome across tissues and individuals. Science (80- ) [Internet]. American Association for the Advancement of Science; 2015 [cited 2020 Oct 29];348:660–5. Available from: https://pubmed.ncbi.nlm.nih.gov/25954002/

37. Heinz S, Romanoski CE, Benner C, Glass CK. The selection and function of cell type-

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

specific enhancers [Internet]. Nat. Rev. Mol. Cell Biol. Nature Publishing Group; 2015 [cited 2020 Oct 29]. p. 144–54. Available from: https://pubmed.ncbi.nlm.nih.gov/25650801/

38. Ko JY, Oh S, Yoo KH. Functional enhancers as master regulators of Tissue-Specific gene regulation and cancer development [Internet]. Mol. Cells. Korean Society for Molecular and Cellular Biology; 2017 [cited 2020 Oct 29]. p. 169–77. Available from: https://pubmed.ncbi.nlm.nih.gov/28359147/

39. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. Nature Publishing Group; 2017;550:204–13.

40. Valiathan R, Ashman M, Asthana D. Effects of Ageing on the Immune System: Infants to Elderly. Scand J Immunol [Internet]. Blackwell Publishing Ltd; 2016 [cited 2020 Oct 29];83:255–66. Available from: https://pubmed.ncbi.nlm.nih.gov/26808160/

41. Johnson ND, Wiener HW, Smith AK, Nishitani S, Absher DM, Arnett DK, et al. Non-linear patterns in age-related DNA methylation may reflect CD4+ T cell differentiation. Epigenetics [Internet]. Taylor and Francis Inc.; 2017 [cited 2020 Oct 29];12:492–503. Available from: https://pubmed.ncbi.nlm.nih.gov/28387568/

42. Tsai PC, Glastonbury CA, Eliot MN, Bollepalli S, Yet I, Castillo-Fernandez JE, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health 06 Biological Sciences 0604 Genetics. Clin Epigenetics [Internet]. BioMed Central Ltd.; 2018 [cited 2020 Oct 29];10. Available from: https://pubmed.ncbi.nlm.nih.gov/30342560/

43. 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 [Internet]. BMJ Publishing Group; 2018 [cited 2020 Oct 30];8. Available from: https://pubmed.ncbi.nlm.nih.gov/30206078/

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

44. 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 [Internet]. Int J Epidemiol; 2013 [cited 2020 Oct 30];42:978–91. Available from: https://pubmed.ncbi.nlm.nih.gov/23064411/

45. Heude B, Forhan A, Slama R, Douhaud L, Bedel S, Saurel-Cubizolles M-JJ, et al. Cohort Profile: The EDEN mother-child cohort on the prenatal and early postnatal determinants of child health and development. Int J Epidemiol [Internet]. Oxford University Press; 2016 [cited 2020 Oct 30];45:353–63. Available from: https://pubmed.ncbi.nlm.nih.gov/26283636/

46. 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 [Internet]. 2012 [cited 2015 Apr 27];41:930–40. Available from: http://ije.oxfordjournals.org/content/41/4/930.long

47. Grazuleviciene R, Danileviciute A, Nadisauskiene R, Vencloviene J. Maternal smoking, GSTM1 and GSTT1 polymorphism and susceptibility to adverse pregnancy outcomes. Int J Environ Res Public Health [Internet]. Int J Environ Res Public Health; 2009 [cited 2020 Oct 30];6:1282–97. Available from: https://pubmed.ncbi.nlm.nih.gov/19440446/

48. 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 [Internet]. Oxford University Press; 2016 [cited 2020 Oct 30];45:382–8. Available from: https://pubmed.ncbi.nlm.nih.gov/27063603/

49. 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 [cited 2020 Oct 30];46:1392-1393k. Available from: https://pubmed.ncbi.nlm.nih.gov/29040580/

50. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al.

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics [Internet]. 2014 [cited 2015 Jan 9];30:1363–9. Available from: http://bioinformatics.oxfordjournals.org/content/30/10/1363

51. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16:37.

52. van Iterson M, Tobi EW, Slieker RC, den Hollander W, Luijk R, Slagboom PE, et al. MethylAid: Visual and interactive quality control of large Illumina 450k data sets. Bioinformatics [Internet]. 2014 [cited 2015 Oct 7];30:3435–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25147358

53. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, et al. Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol [Internet]. BioMed Central Ltd.; 2014 [cited 2020 Oct 30];15. Available from: https://pubmed.ncbi.nlm.nih.gov/25599564/

54. Fortin J-P, Fertig E, Hansen K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. F1000Research [Internet]. 2014 [cited 2015 Oct 7];3:175. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4176427&tool=pmcentrez&rendertype=abstract

55. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics [Internet]. 2007 [cited 2014 Jul 10];8:118–27. Available from: http://biostatistics.oxfordjournals.org/content/8/1/118.abstract

56. Buckberry S, Bent SJ, Bianco-Miotto T, Roberts CT. MassiR: A method for predicting the sex of samples in gene expression microarray datasets. Bioinformatics [Internet]. Oxford University Press; 2014 [cited 2020 Oct 30];30:2084–5. Available from:

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

https://pubmed.ncbi.nlm.nih.gov/24659105/

57. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet [Internet]. PLoS Genet; 2007 [cited 2020 Oct 30];3:1724–35. Available from: https://pubmed.ncbi.nlm.nih.gov/17907809/

58. Houseman EAE, Accomando WP, Koestler DDC, Christensen BBC, Marsit CCJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinformatics [Internet]. 2012 [cited 2015 May 12];13:86. Available from: http://www.biomedcentral.com/1471-2105/13/86

59. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet [Internet]. 2007 [cited 2017 Oct 20];81:559–75. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17701901

60. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience [Internet]. 2015 [cited 2018 Feb 23];4:7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25722852

61. Pedersen BS, Quinlan AR. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Peddy. Am J Hum Genet [Internet]. Elsevier; 2017 [cited 2018 Feb 2];100:406–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28190455

62. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature [Internet]. 2015 [cited 2017 Sep 26];526:68–74. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26432245

63. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet [Internet]. 2016 [cited 2017 May

bioRxiv preprint doi: https://doi.org/10.1101/2020.11.05.368076; this version posted November 5, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29];48:1284–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27571263

64. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet [Internet]. 2016 [cited 2017 May 29];48:1279–83. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27548312

65. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Reshef YA, Finucane HK, et al. Reference-based phasing using the Haplotype Reference Consortium panel. Nat Genet [Internet]. Nature Publishing Group; 2016 [cited 2020 Oct 30];48:1443–8. Available from: https://pubmed.ncbi.nlm.nih.gov/27694958/

66. Fuchsberger C, Abecasis GR, Hinds DA. Minimac2: Faster genotype imputation. Bioinformatics [Internet]. Oxford University Press; 2015 [cited 2020 Oct 30];31:782–4. Available from: https://pubmed.ncbi.nlm.nih.gov/25338720/

67. J AA and R. topGO: topGO: Enrichment analysis for Gene Ontology. No Title. 2010. p. R package version 2.22.0.

68. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics [Internet]. Bioinformatics; 2012 [cited 2020 Nov 2];28:1353–8. Available from: https://pubmed.ncbi.nlm.nih.gov/22492648/