

METHODOLOGY

Open Access



Unified high-resolution immune cell fraction estimation in blood tissue from birth to old age

Xiaolong Guo^{1†}, Mahnoor Sulaiman^{2,3†}, Alexander Neumann³, Shijie C. Zheng⁴, Charlotte A.M. Cecil^{3,5*}, Andrew E. Teschendorff^{1*} and Bastiaan T. Heijmans^{2*}

Abstract

Variations in immune-cell fractions can confound or hamper interpretation of DNAm-based biomarkers in blood. Although cell-type deconvolution can address this challenge for cord and adult blood, currently there is no method applicable to blood from other age groups, including infants and children. Here we construct and extensively validate a DNAm reference panel, called UniLIFE, for 19 immune cell-types, applicable to blood tissue of any age. We use UniLIFE to delineate the dynamics of immune-cell fractions from birth to old age, and to infer disease associated immune cell fraction variations in newborns, infants, children and adults. In a prospective longitudinal study of type-1 diabetes in infants and children, UniLIFE identifies differentially methylated positions that precede type-1 diabetes diagnosis and that map to diabetes related signaling pathways. In summary, UniLIFE will improve the identification and interpretation of blood-based DNAm biomarkers for any age group, but specially for longitudinal studies that include infants and children. The UniLIFE panel and algorithms to estimate cell-type fractions are available from our EpiDISH Bioconductor R-package: <https://bioconductor.org/packages/release/bioc/html/EpiDISH.html>

Keywords Immune system, Cord blood, DNA methylation, EWAS, Cell-type deconvolution

Background

DNA methylation (DNAm) is an epigenetic mark that plays an important role in gene regulation, aging and disease [1]. Like gene expression, DNAm differs significantly between cell-types [2–5]. Due to the high cost and time consuming task of cell-sorting, especially when implemented at large scale (e.g. as in the case of population-based studies), epigenome studies typically examine DNAm in bulk tissue, which comprises a mixture of many different cell-types [1]. Thus, DNAm measurements in bulk-tissue are subject to potential confounding by variations in cell-type composition [6–8]. To address this challenge, cell-type deconvolution algorithms and DNAm reference panels have been developed [9–13] that allow researchers to estimate cell-type fractions, which are necessary to adjust for the underlying cell-type heterogeneity, thus circumventing the need to experimentally measure the actual cell-type composition.

[†]Xiaolong Guo and Mahnoor Sulaiman contributed equally to this work.

*Correspondence:

Charlotte A.M. Cecil
c.cecil@erasmusmc.nl
Andrew E. Teschendorff
andrew@sinh.ac.cn
Bastiaan T. Heijmans
b.t.heijmans@lumc.nl

¹ Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China

² Department of Biomedical Data Sciences, Leiden University Medical Center, Einthovenweg 20, Leiden 2333 ZC, The Netherlands

³ Department of Child and Adolescent Psychiatry/Psychology, Sophia's Children Centre, Erasmus MC, Rotterdam, The Netherlands

⁴ Pfizer Research & Development, Pfizer Inc, Groton, CT, USA

⁵ Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands



One of the most commonly used bulk tissues for epigenetic studies is blood, including cord blood at birth [14–21] and whole/peripheral blood at later ages [22]. While both involve ‘blood’, there are key differences between these tissue types, with cord blood containing a unique set of cell types that rapidly decline after birth, e.g. nucleated red blood cells (nRBC), as well as cell-types that have different properties compared to their adult counterparts, for example the degree of ‘naivety’ of B or T cells [23]. To capture these differences, distinct cell-type estimation panels, called DNAm reference matrices/panels, have been created [10, 24–28]. The Gervin DNAm reference matrix, which estimates fractions for seven major cell groups in cord-blood (B cells, Granulocytes, Natural Killer (NK) cells, Monocytes, CD4 T, CD8 T, and nRBCs), is considered the gold standard for studies using DNAm data at birth [25]. For adult blood, the Salas [27] and updated EpiDISH [29] DNAm reference matrices estimate fractions for 12 immune cell types. While both types of panels are effective within their specific contexts, i.e. cord-blood/birth and adulthood, respectively, the use of separate DNAm reference panels for different life stages can pose key challenges for the field. In particular, any analysis aiming to investigate longitudinal dynamics or epigenetic timing effects using blood at different time points, say cord-blood and mid-age, would need to apply different cell-type estimation panels, which could cause confounding between tissue composition and epigenetic changes that happen in individual cell-types [30]. Moreover, current cord and adult-blood reference panels are not directly applicable to blood sampled in between these ages, say from infants, children or young adolescents, since in the early life stages, immune cells are undergoing maturation with the corresponding fractions also displaying highly variable dynamics [23, 31–33]. As such, the development of a ‘universal’ or ‘lifecourse’ blood DNAm reference matrix would enable researchers to use the same panel at any time point, thus reducing confounding between timepoints, and enhancing the reliability and interpretability of DNAm-based biomarkers [34, 35]. Given the increasing number of cohort studies with DNAm measured repeatedly across time, specially in infants, children and adolescents [31, 36–39], there is an urgent need to develop such a lifecourse DNAm reference panel.

The purpose of this study is to create and extensively validate a novel, unified ‘lifecourse’ DNAm reference panel, called UniLIFE, that can be applied to blood samples of any age, from birth onwards, including infancy, childhood, adolescence and adulthood. Our validation is performed on (i) independent sorted samples from cord and adult blood derived with two different technologies (Illumina beadarrays and whole-genome bisulfite

sequencing (WGBS)), (ii) on unsorted cord and adult blood samples with matched cell-counts, and (iii) on blood samples from infants, children and adolescents. We use our lifecourse blood DNAm reference panel to delineate the dynamic changes in blood composition that occur immediately after birth and during the first months and years of life, a time period for which current panels are not directly applicable and where we also currently lack much knowledge [40]. Indeed, charting the dynamic landscape of immune-cell composition in blood during the earliest life-stages is critically important as this can also shed light on future disease predisposition or indicate compositional changes that are diagnostic of disease [40, 41]. In line with this, we demonstrate that our reference panel can capture disease-associated cell-type composition variations during the first years of life, including DNAm changes that predate the onset of type-1 diabetes and which map to relevant signaling pathways.

Methods

A brief summary of all main DNAm datasets used here and their purpose is provided in Additional File 1: Table S1. Below we describe the DNAm datasets used for reference construction, technical and biological validations.

Processing of the cord and adult blood DNAm datasets used in reference construction

Human sorted cord blood DNA methylation datasets used for building reference

Bakulski [42]: The HM450k dataset was downloaded from the Bioconductor package *FlowSorted.CordBlood*. 450 k, which provides a raw *RGchannelSet* object containing all sample information. In R (version 4.4.0), the *preprocessRaw()* and *ratioConvert()* functions from the *minfi* package (version 1.49.1) were used to convert the Red and Green channels of this object into a matrix of Beta values. The *getAnnotation()* function of the *minfi* package was used to identify and remove 11,458 probes located on sex chromosomes, and the *dropLociWithSnps()* function was employed to exclude 17,541 probes with potential SNPs in the CpG site and SNPs in the single base extension site. Additionally, 26,569 probes identified as having cross-reactivity were also excluded. A total of 1,787 probes were excluded for failing overall detection with *P*-values > 0.01 in any sample. The BMIQ algorithm was applied to scale bias of Type 2 probes. The final analysis included data on 428,157 probes across 104 samples, including 15 whole blood, 12 granulocytes, 15 monocytes, 15 B cells, 14 NK cells, 15 CD4 + T cells, 14 CD8 + T cells, and 4 nRBCs (nucleated Red Blood Cells).

Gervin [25]: The HM450k dataset was downloaded from the R package *FlowSorted.CordBlood-Norway.450*

k in Bioconductor, which also provides a *RGchannelSet* object. Sample information such as cell type and gender are also contained within this object. The same preprocessing pipeline was applied as above, which involved the exclusion of 17,541 SNP probes, 11,458 sex chromosome probes, 26,569 cross-reactive probes, and 1,066 probes with a *P*-value > 0.01 in any sample. Ultimately, this resulted in data for 428,878 probes across 77 samples (with each of the following cell types has 11 samples: B cells, CD4 + T cells, CD8 + T cells, granulocytes, monocytes, NK cells, and white blood cells). Type-2 bias was corrected using BMIQ. Notably, this dataset does not possess nRBCs.

de Goede_1 [43, 44]: The HM450k dataset was downloaded from the GEO data repository (GEO accession GSE68456). After reading the IDAT files into a *RGchannelSet* object using the *read.metharray.exp()* function from the *minfi* package, the same preprocessing pipeline was applied as mentioned above, which involved the exclusion of 17,541 SNP probes, 11,458 sex chromosome probes, 26,569 cross-reactive probes, and 2,070 probes with a *P*-value > 0.01 in any sample. Finally, this resulted in data for 427,874 probes across 62 samples. Type-2 bias was corrected using BMIQ. The authors employed two sorting protocols: "standard" and "stringent." Whole (CD3 +) T cells, monocytes, and nRBCs were collected from five individuals using the standard sorting method; B cells, CD4 T cells, CD8 T cells, granulocytes, monocytes, NK cells, and nRBCs were obtained from seven individuals following the stringent sorting protocol. In total, there were 7 B cells, 7 CD4 + T cells, 6 CD8 + T cells, 7 granulocytes, 12 monocytes (5 from standard and 7 from stringent), 6 NK cells, 5 nRBCs (from the standard protocol), and 7 nRBCs (from the stringent protocol), as well as 5 T cells. They have demonstrated that the standard sorting protocol can lead to cross-contamination among T cells, monocytes, and nRBCs. Therefore, subsequent analyses should exclude these 15 samples obtained from the standard protocol and instead use only those acquired through the stringent sorting protocol.

Lin [45]: The EPIC dataset was separated from the R package *FlowSorted.CordBloodCombined.450 k* in Bioconductor, as the original source was inaccessible. Within the obtained *RGchannelSet* object, Gervin et al. retained only the HM450k probes. Then the aforementioned preprocessing pipeline was applied, which involved the exclusion of 16,353 SNP probes, 10,446 sex chromosome probes, 24,823 cross-reactive probes, and 918 probes with a *P*-value > 0.01 in any sample. The final analysis included data on 400,553 probes across 83 samples, which comprised 14 granulocytes, 14 monocytes, 13 B cells, 14 NK cells, 14 CD4 + T cells, and 14 CD8 + T cells.

Type-2 bias was corrected using BMIQ. This dataset does not possess nRBCs too.

Processing of the cord and adult blood DNAm datasets used for validation

Human cord blood DNA methylation datasets used for validation

De Goede_2 [46, 47]: The HM450k sorted cord blood dataset was downloaded from the GEO data repository (GEO accession GSE82084). 36 samples in total: 5 each of T cells, monocytes, granulocytes and nRBCs from term births; and 5 T cells, 4 monocytes, 4 nRBCs and 3 granulocytes from preterm births. *Minfi* package was used to preprocess the raw IDAT files. Subsequently, we excluded 17,541 SNP probes, 11,458 sex chromosome probes, 26,569 cross-reactive probes, and 598 probes with a *P*-value greater than 0.01 in any sample. This resulted in data for 426,642 probes. Type-2 bias was corrected using BMIQ.

Salas [27, 48]: The EPIC dataset was downloaded from the GEO data repository (GEO accession GSE180970). There are 12 umbilical cord blood artificial mixtures with six cell types, including CD8 T, CD4 T, NK, B cell, monocyte, and granulocyte. After obtaining the raw IDAT files, 790,868 probes remain under the same preprocessing workflow. Type-2 bias was corrected using BMIQ.

Jones [25, 49]: The HM450k cord blood dataset was downloaded from the GEO data repository (GEO accession GSE127824). The dataset includes 24 human whole cord blood samples with matched FACS counts of seven cell types, including CD8 T, CD4 T, NK, B cell, nRBC, monocyte, and granulocyte. *Minfi* package was used to preprocess the raw IDAT files and there are 429,346 probes remained. Type-2 bias was corrected using BMIQ.

Hernandez [50, 51]: The HM450k cord blood dataset was downloaded from the GEO data repository (GEO accession GSE149572). The dataset includes 3 adult human blood samples and 20 human cord blood samples. Beta value matrix with *p*-values was downloaded, and after removing the probes described earlier, 428,055 probes remain. Type-2 bias was corrected using BMIQ. Three adult samples were excluded from the subsequent analysis.

Puvvula [52, 53]: The HM450k cord blood dataset was downloaded from the GEO data repository (GEO accession GSE269983). The dataset includes 55 human whole cord blood mononuclear cell samples, 68 fetal side placenta tissues and 69 maternal side placenta tissues. *Minfi* package was used to preprocess the raw IDAT files and there are 422,967 probes remained. Type-2 bias was corrected using BMIQ. Only 55 cord blood samples were used for subsequent analysis.

Holland [54, 55]: The HM450k cord blood dataset was downloaded from the GEO data repository (GEO accession GSE97628). The dataset includes 380 human whole cord blood samples. Minfi package was used to preprocess the raw IDAT files and there are 391,442 probes remained. Type-2 bias was corrected using BMIQ.

Bell_1 [56, 57]: The HM450k cord blood dataset was downloaded from the GEO data repository (GEO accession GSE188949). This dataset includes 107 cord blood samples from preterm neonates, of which 14 samples are from neonates with Bronchopulmonary Dysplasia (BPD). Minfi package was used to preprocess the raw IDAT files and there are 429,730 probes remained. Type-2 bias was corrected using BMIQ. Fourteen samples with BPD were excluded from the subsequent analysis.

Peripheral blood datasets of children used for validation

Barwick [58, 59]: The HM450k peripheral blood leukocytes dataset from healthy children was downloaded from the GEO data repository (GEO accession GSE36064). This dataset contains 78 samples from children aged 1 to 16 years. Minfi package was used to preprocess the raw IDAT files and there are 428,593 probes remained. Type-2 bias was corrected using BMIQ.

Quinn [60, 61]: The EPIC venous blood dataset was downloaded from the GEO data repository (GEO accession GSE224573). Sample size at each timepoint was as follows: $n = 64$ at birth, $n = 8$ at six months, $n = 35$ at one year, $n = 32$ at two years, and $n = 16$ at three years. Minfi package was used to preprocess the raw IDAT files and there remained 790,585 probes after QC. Type-2 bias was corrected using BMIQ.

Wang [62, 63]: The HM450k dataset was downloaded from the GEO data repository (GEO accession GSE104812). Samples included 29 boys and 19 girls, and their ages ranged from 6 to 14 years. Minfi package was used to preprocess the raw IDAT files and there remained 429,426 probes after QC. Type-2 bias was corrected using BMIQ.

Human adult whole and sorted blood DNA methylation datasets used for validation

Bell_2 [64, 65]: The raw IDAT files for 789 sorted immune-cell samples of the HM450k platform were downloaded from the GEO data repository (GEO accession GSE224807). Any probe with a P -value greater than 0.01 was excluded as well as SNP probes and sex chromosome probes, resulting in a total remaining count of 456,116 probes. Type-2 bias was corrected using BMIQ. Among these 789 samples, there are six leukocyte subtypes: monocyte, granulocyte, B cell, CD4 + T cell, CD8 + T cell, and natural killer cell, as well as 36 whole blood samples (removed in subsequent analysis).

Other adult blood datasets used for validation have been previously normalized and analyzed in Luo et al., including: (1) *GSE112618* [66]: 6 EPIC whole blood samples with matched FACS counts of 7 adult immune cell types, including CD4 T cell, CD8 T cell, B cell, NK, Monocyte, Neutrophil, and Eosinophil; (2) *GSE167998* [67]: 12 EPIC artificial mixtures composed of 10 adult immune cell types: naïve B cell, memory B cell, naïve CD4 T cell, memory CD4 T cell, CD8 T cell, NK, Mono, Neu, Eos, and Regulatory T cell (Treg); (3) *GSE56581* [68]: 1202 monocyte and 214 CD4 + T-cell 450 k samples. (4) *BLUEPRINT* [69]: 139 monocyte, 139 naïve CD4 + T-cell, and 139 neutrophil 450 k samples from the same 139 individuals. (5) *E-MTAB-2145* [70]: 6 CD4 + T-cell, 6 CD8 + T-cell, 6 B-cell, 6 neutrophil, and 6 monocyte 450 k samples. (6) *GSE65097* [71]: 15 neutrophil 450 k samples. (7) *GSE50222* [72]: 8 CD4 + T-cell 450 k samples. (8) *GSE63499* [73]: 12 neutrophil 450 k samples. (9) *GSE71955* [74]: 31 CD4 + T-cell and 31 CD8 + T-cell 450 k samples. (10) *GSE71244* [75]: 6 CD4 + T-cell, 5 CD8 + T-cell, 4 B-cell, and 5 monocyte 450 k samples. (11) *GSE59250* [76]: 71 CD4 + T-cell, 56 B-cell, and 28 monocyte 450 k samples. (12) *GSE59065* [77]: 99 CD4 + T-cell and 100 CD8 + T-cell 450 k samples. (13) *EGAS00001001598* [78]: 49 CD4 + T-cell, 50 B-cell, and 52 monocyte 450 k samples. (14) *GSE35069* [79]: 6 CD4 + T-cell, 6 CD8 + T-cell, 6 B-cell, 6 neutrophil, 6 monocyte, and 6 eosinophil 450 k samples. (15) HPT-450 k (GSE210254) [80]: 418 HM450k PB samples; (16) Zannas (GSE72680) [81]: 422 HM450k PB samples; (17) Johansson (GSE87571) [82]: 732 HM450k PB samples; (18) HNM (GSE40279) [83]: 656 HM450k PB samples.

Processing and analysis of sorted immune cell WGBS data from IHEC

From the IHEC data portal (<https://epigenomesportal.ca/ihec/>), we obtained WGBS hg38 immune-cell-sorted samples from build version 2020–10 [84]. This included the following adult immune cell-types: 4 memory CD4 + T-cell, 2 T-regulatory, 2 naïve CD8 + T-cell, 2 memory CD8 + T-cell, 5 naïve B-cell, 5 memory B-cell, 10 neutrophil, 20 monocyte, 2 eosinophil and 2 natural killer cell samples. It also included the following cord-blood subtypes: 2 cord blood CD8 T, 2 cord blood B-cell, 2 cord blood neutrophil, 2 cord blood monocyte, and 2 cord blood natural killer. The downloaded files were in bigwig format, and processed as described by us previously [29]. For each WGBS sample, we computed the overlap with UniLIFE probes. An average of 1560 UniLIFE CpGs were covered by at least 20 reads. To estimate purity of these samples, UniLIFE was applied to infer the underlying fractions. For the in-silico mixture analysis, one needs to compute the overall coverage of UniLIFE probes across

all samples used in the mixing. Because this global coverage was too low, for the in-silico mixture analyses we did not impose any threshold on read coverage. Although in any given sample, this relaxation on read-coverage can lead to fluctuations in the beta-values, we note that reference-based cell-type deconvolution can tolerate even up to 30% binary errors (i.e. 0 s/1 s being incorrectly imputed as 1 or 0, respectively) [47].

DNA methylation datasets used for biological validation

Lorente-Pozo et al. [85, 86]: This EPIC whole blood dataset was downloaded from the GEO data repository (GEO accession GSE155952). Blood samples were from septic ($n = 17$) and healthy ($n = 6$) preterm infants. Minfi package was used to preprocess the raw IDAT files. A total of 789,721 probes remained. Type-2 bias was corrected using BMIQ.

Henneman et al. [87, 88]: This HM450k whole blood dataset was downloaded from the GEO data repository (GEO accession GSE107211). Blood samples were obtained from 10 newborns with Down syndrome and 5 age-matched non-trisomic newborns. Minfi package was used to preprocess the raw IDAT files. A total of 429,322 probes remained. Type-2 bias was corrected using BMIQ.

Naumova et al. [89, 90]: This HM450k whole blood dataset was downloaded from the GEO data repository (GEO accession GSE174555). Blood samples were obtained from 17 children with Down syndrome and 17 typically developing children (age ranges from 6 to 56 months). Minfi package was used to preprocess the raw IDAT files. A total of 417,613 probes remained. Type-2 bias was corrected using BMIQ.

Bacalini et al. [91, 92]: This HM450k whole blood dataset was downloaded from the GEO data repository (GEO accession GSE52588). Whole blood samples were obtained from 29 Down syndrome patients (age ranges from 10 to 43 years) and 58 healthy controls (age ranges from 9 to 83 years). Minfi package was used to preprocess the raw IDAT files. A total of 409,225 probes remained. Type-2 bias was corrected using BMIQ. To ensure age matching, subsequent analyses only included 24 Down syndrome samples and 31 healthy control samples, all between the ages of 18 and 50.

Chen et al. [93, 94]: This HM450k whole blood dataset was downloaded from the GEO data repository (GEO accession GSE152084). Peripheral blood samples were taken from a total of 48 individuals (24 AD patients and 24 controls, age ranges from 0.1 year to 5.4 years). Minfi package was used to preprocess the raw IDAT files. A total of 415,890 probes remained. Type-2 bias was corrected using BMIQ.

Yang et al. [95, 96]: This HM450k peripheral blood mononuclear cell (PBMC) dataset was downloaded from

the GEO data repository (GEO accession GSE40576). PBMCs from 97 atopic asthmatic and 97 nonatopic non-asthmatic children (age ranges from 6 to 12 years). Minfi package was used to preprocess the raw IDAT files. A total of 428,167 probes remained. Type-2 bias was corrected using BMIQ.

Langie et al. [97, 98]: This HM450k PBMC dataset was downloaded from the GEO data repository (GEO accession GSE110128). It contains 29 Respiratory Allergy samples and 61 healthy control samples from children approximately 10 years old. Minfi package was used to preprocess the raw IDAT files. A total of 427,883 probes remained. Type-2 bias was corrected using BMIQ.

Processing of DNAm datasets of type-1 diabetes

DAISY study [99, 100]: The Illumina DNAm dataset was downloaded from the GEO data repository (GEO accession GSE142512). This dataset contains 395 samples from 174 individuals. The ages of the individuals range from 6 months to 22 years, with the majority being under 10 years old. A total of 211 samples were sequenced using the EPIC array, and 184 samples were sequenced using the 450 k array. Half of the individuals were eventually diagnosed with type 1 diabetes, but all samples were taken before diagnosis. Samples from different array types were preprocessed separately. In R (version 4.4.0), the `preprocessRaw()` and `ratioConvert()` functions from the minfi package (version 1.49.1) were used to convert the Red and Green channels of IDAT files into a matrix of Beta values. Considering that the authors removed 7 arrays with inconsistent gender records during the quality checks, the `estimateSex` function from the R package `wateRmelon` (version = 1.0) was used to test the consistency between estimated and recorded gender. It was confirmed that the 7 arrays had already been removed from the raw IDAT data. The `getAnnotation()` function of the minfi package was used to identify and remove probes located on sex chromosomes (450 k = 11,458, EPIC = 19,298), and the `dropLociWithSnps()` function was employed to exclude probes with potential SNPs in the CpG site and SNPs in the single base extension site (450 k = 17,541, EPIC = 30,345). Additionally, probes identified as having cross-reactivity were also excluded (450 k = 26,569, EPIC = 39,605). Any beta value with a detection *P*-value greater than 0.01 was considered as 'missing', and any probe with a coverage-fraction less than 0.9 was discarded. Missing values were then imputed using `impute.knn` ($k = 5$) of R package `impute` (version = 1.72.3). Type-2 bias was corrected using BMIQ algorithm. Finally, the EPIC and 450 k normalized datasets were merged, defining a DNAm matrix over 400,485 common probes and 395 samples derived from 174 subjects. Of these 395 samples, 196 were from 87 subjects

who would later develop T1D, with 199 samples being derived from 87 age-matched controls. Both controls and prospective T1D cases were sampled multiple times, with 22 subjects sampled 4 times, 49 subjects sampled 3 times, 57 subjects sampled twice and 46 subjects sampled only once. Samples were collected at 4 timepoints, two timepoints before the diagnosis of islet autoimmunity (IA), and two timepoints after diagnosis of IA but before diagnosis of T1D. Age range was 6 months to 22 years.

Paul et al. [78, 101]: This is an Illumina 450 k DNAm dataset of sorted monocytes, B-cells and T-cells from monozygotic twin pairs discordant for T1D, available from European Genome-phenome Archive (EGA) with the accession code EGAS00001001598 (<https://www.ebi.ac.uk/ega/studies/EGAS00001001598>). We used the normalized DNAm dataset as described in the original publication [101], consisting of 3 separate DNAm data matrices defined over 406,365 probes and 100 B-cell samples (50 twin pairs), 98 T-cell samples (49 twin pairs) and 104 monocyte samples (52 twin pairs).

Construction of the universal 19 immune cell-type DNAm reference matrix

First, we focused on Illumina DNAm datasets profiling sorted cord blood cell-types that include CD4 + T-cells (CD4 T), CD8 + T-cells (CD8 T), NK-cells, B-cells, Monocytes (Mono), Granulocytes (Gran) and nucleated red blood cells (nRBCs). In total, we obtained four cord-blood Illumina 450 k DNAm studies: Goede [43], Gervin [25], Bakulski [42] and Lin [45]. All 4 studies contained on the order of 10 samples for each of the lymphoid and myeloid subsets, but only Bakulski and Goede contributed nRBCs ($n = 4$ and $n = 7$, respectively). For quality control, we used hierarchical clustering over the top significant singular vectors to cluster samples in each study. In three studies (Gervin, Lin and Goede), samples clustered perfectly by cell-type, with the exception of intermixing of CD4 T and CD8 T cells in Gervin. We verified using supervised analysis that we could identify many significant DMCs with large fold changes between the CD4 T and CD8 T cells from Gervin, suggesting that the observed intermixing from the unsupervised analysis is not necessarily a reflection of low purity. In the remaining study (Bakulski), nRBCs clustered separately from all other cell-types, but these sometimes did not segregate perfectly from each other, suggesting low purity of some of the leukocyte subsets. Hence, from Bakulski et al., we decided to only take forward the nRBC subset. We then merged the normalized DNAm data from the 4 studies, and performed a SVD and hierarchical clustering over the significant components. Whilst samples clustered by cell-type according to the first two singular vectors, the hierarchical

clustering revealed some batch effects, for instance, the NK-cells of Goede clustered separately from the NK-cells of Lin and Gervin. Based on this hierarchical clustering we thus removed samples that did not cluster as expected. We subsequently reapplied SVD and hierarchical clustering, which resulted in a clustering where samples segregated perfectly by cell-type, with the exception of NK and B-cells. However, this splitting up of NK and B-cells had not been observed in our previous HC iteration, suggesting that this is a more minor effect. We argued that these residual batch effects are small and would not affect our downstream supervised analysis where we identify DMCs between cell-types. Supporting this, a scatterplot of the merged samples over the top two singular vectors resulted in a perfect segregation of samples by cell-type. The resulting sample size distribution for each cell-type is as follows: 28 B-cells (8 Gervin, 7 Goede, 13 Lin), 29 CD4 T-cells (8 Gervin, 7 Lin, 14 Lin), 20 CD8 T-cells (6 Goede, 14 Lin), 24 Granulocytes (8 Gervin, 2 Goede, 14 Lin), 32 Monocytes (11 Gervin, 7 Goede, 14 Lin), 22 NK-cells (8 Gervin, 14 Lin), 11 nRBCs (4 Bakulski, 7 Goede). Thus, for each cell-type, samples are drawn from at least 2 cohorts, and these samples cluster together, suggesting that their DNAm patterns are robust to residual batch effects. Overall, our merged sorted cord blood DNAm dataset was defined over 397,217 CpG probes and 166 samples encompassing 7 cord blood cell-types.

Next, we integrated this merged cord blood DNAm dataset with the EPIC DNAm dataset from Salas et al. [27] encompassing 54 samples from 12 adult immune cell-types, a dataset which we have previously normalized and analysed in Luo et al. [29]. To distinguish these adult immune cell-types from those in cord-blood, we add the prefix “a” to indicate they are from adults: aBnv (adult naïve B-cells), aBmem (adult memory B-cells), aCD4 Tnv (adult naïve CD4 T-cells), aCD4 Tmem (adult memory CD4 T-cells), aCD8 Tnv (adult naïve CD8 T-cells), aCD8 Tmem (adult memory CD8 + T-cells), aTreg (adult T-regulatory cells), aNK (adult NK-cells), aNeu (adult Neutrophils), aMono (adult Monocytes), aEos (adult Eosinophils), aBaso (adult Basophils). Thus, this resulted in an overall integrated DNAm dataset defined over 375,506 CpGs and 220 sorted samples, encompassing 19 immune cell-types (7 from cord blood, 12 from adult blood). We performed a SVD/PCA on this overall integrated DNAm dataset to ensure that the top PCs/SVs capture biology instead of batch effects.

To build the DNAm reference matrix for 19 immune cell-types, we need to identify cell-type specific DNAm markers. To identify these cell-type specific markers we used the following procedure:

1. For each cell-type, we performed differential DNAm analysis using an empirical Bayes method (limma) [102, 103], comparing the cell-type to all the rest, using a stringent FDR < 0.001 threshold.
2. Demanding for (i) hypomethylated cell-type specific markers that the attained maximum value over samples of that cell-type is less than the minimum value across all other cell-types. Likewise for (ii) hypermethylated markers we demand that the minimum value for the cell-type is greater than the maximum over all other cell-types.

We note that the second criterion is critically important to ensure specificity for a given cell-type, because criterion-1 alone does not guarantee specificity for each cell-type. For each cell-type, we finally ranked (i) hypomethylated CpGs by the gap in DNAm between the maximum value attained in the cell-type of interest and the minimum value across all other cell-types, and (ii) hypermethylated CpGs by the gap in DNAm between the minimum value attained in the cell-type of interest and the maximum value across all other cell-types. Thus, for each cell-type, we obtained a number of cell-type specific hypomethylated and hypermethylated markers. Aiming for approximately 100 markers per cell-type, we next tailored the marker selection for each cell-type as follows: for aBaso, aCD4 Tmem, aBmem, aTreg, aCD8 Tmem, aEos, aNK, aNeu, and aMono, we selected the top 100 hypomethylated marker CpGs, since the number of valid hypomethylated marker CpGs was greater than 100 for these adult cell-types. For the adult naïve fractions aCD4 Tnv and aBnv we selected the top 50 hypomethylated and the top 50 hypermethylated markers. For aCD8 Tnv, there were insufficient numbers of hypermethylated markers, so we selected the top 82 hypomethylated and the top 18 hypermethylated ones. For the cord-blood cell-types, the numbers of hypomethylated markers was generally always less than 100, requiring an approach that balanced hypo and hypermethylated markers. Specifically, for B-cells and nRBCs, we were able to choose the top 50 hypo and top 50 hypermethylated markers. For the remaining 5 cord blood cell-types (CD4 T, CD8 T, Mono, Gran and NK), the total number of hypo and hypermethylated markers was less than 100. Hence, for these cell-types we decided to supplement the marker CpGs with the top-100 hypomethylated CpGs ranked by fold-change, and not by the specificity gap score. Finally, we required marker CpGs to also display a range of DNAm-values across all cell-types of at least 0.2. In total, this resulted in 1906 marker CpGs, with approximately 100 markers per cell-type. The final DNAm reference matrix defined over these 1906 cell-type specific marker CpGs and 19

immune-cell types was built by taken the median value over the samples of a given cell-type.

Estimation of cell-type fractions

In all cases, given the 19 cell-type DNAm reference matrix for either the Illumina EPIC or 450 k dataset, we estimated corresponding cell-type fractions using the EpiDISH Bioconductor R-package. Specifically, we ran the *epidish* function with “RPC” as the method and *maxit* = 500.

In-silico mixture validation analysis

In our study, we constructed three types of in-silico mixtures using the sorted cord blood dataset and the sorted adult blood dataset to simulate whole cord blood, whole adult blood, and infant whole blood respectively. The simulation of cord blood utilized all samples from de Goede_2, which include 10 T-cells, 9 monocytes, 9 nRBCs, and 8 granulocytes. For the simulation of adult whole blood, we utilized adult samples from the IHEC WGBS dataset. Therefore, the in-silico mixtures were generated from 4 memory CD4 + T-cell, 2 T-regulatory, 2 naïve CD8 + T-cell, 2 memory CD8 + T-cell, 5 naïve B-cell, 5 memory B-cell, 10 neutrophils, 20 monocytes, 2 eosinophils, and 2 natural killer samples. To simulate infant whole blood, we combined sorted cord blood samples and sorted adult blood samples, and concurrently constructed Illumina 450 k mixtures and WGBS mixtures. The construction of Illumina 450 k in-silico mixtures used all samples from de Goede_2 and Bell datasets, while the WGBS mixtures used all samples from the IHEC dataset (sorted cord blood samples included 2 CD8 T, 2 NK, 2 B cells, 2 Neu, and 2 Monocytes). For each type of mixture, we generated 100 samples, randomly selecting one sample from each immune-cell type and using random weights drawn from a uniform distribution to generate the linear combination. Performance was assessed using Pearson R-values and Root Mean Square Error (RMSE).

Comparison to Generation R blood cell counts

Percentiles of blood cell counts for Generation R study have been published in table E2 of the respective publication [31]. Specifically, for each age group (0 months, 6 m, 14 m, 25 m and 72 m), leukocyte count values for all major immune cell types were given for the 5%, 25%, 50%, 75% and 95% centiles, allowing us to generate comparative boxplots.

Identification of T1D-DMCs in DAISY and monozygotic twin studies

Because the DAISY DNAm dataset was generated with two separate array-types (450 k & EPIC), and PC1 from

a PCA analysis correlated most strongly with array-type, it was necessary to adjust for array-type when searching for T1D-DMCs. In addition, we observed non-negligible technical variation associated with array position and chip. Although cases and controls were correctly randomized and matched for age, timepoint, sex, array-type, position and chip, we adjusted for all of these factors when searching for T1D-DMCs. Specifically, we ran a linear model with DNAm beta-values as the response variable against prospective disease status, adjusting for array-type, array position, chip, sex, age and timepoint. Of note, this selects for DMCs where differential DNAm predates development of T1D in a manner that is also independent of timepoint, i.e. whether samples were collected before or after IA-diagnosis. T1D-DMCs were declared at a significance level threshold of $FDR < 0.05$. To identify CpGs where the differential DNAm between cases and controls increases or decreases with timepoint, we ran a model with an interaction term between disease status and timepoint, which however did not result in any significant hits. In the case of the monozygotic twin study, for each of the 3 cell-types (monocytes, B-cells and T-cells), we ran paired t-tests to estimate t-statistics and *P*-values for each CpG in association with T1D-status.

Case-control prediction comparison of DNAm reference panels

Using the 19 cell-type fractions as estimated with UniLIFE, we applied an elastic net logistic regression model to each of the atopic asthma, respiratory allergy and Down syndrome (DS) datasets so as to predict case-control status. In the case of DS, in order to increase sample size, we combined the datasets profiling neonates, toddlers, and adults. For each dataset, we randomly split the data into 80% training and 20% test sets. The *cv.glmnet* function from the *glmnet* R package was used to train the model, with the mixing parameter alpha set to 0.5, and the optimal regularization parameter lambda selected via tenfold cross-validation. Predictions on the test set were generated using the *predict* function, and the area under the receiver operating characteristic curve (AUC) was calculated with the *roc* and *auc* functions from the *pROC* R package. To ensure robustness, we repeated the training-test set split 1000 times. Using the exact same 1000 training-test set splits, we also trained elastic net regression models from the 12 immune cell-type fractions, as estimated with the 12 adult immune-cell type reference panel from Luo et al. [29]. The 1000 AUC values obtained by UniLIFE and the 12 cell-type panel were then compared using a paired two-tailed Wilcoxon rank sum test.

In a separate analyses, we employed a forward selection strategy for each dataset (atopic asthma, respiratory allergy and merged DS) to build logistic predictors

of case-control status. We first fitted univariate logistic regression models for each cell type using the *glm* function from the *stats* R package, ranking cell types by their coefficient *P*-values. Starting with an intercept-only model, we iteratively added cell types from the ranked order, incorporating each one into a multivariate logistic regression model via *glm*, assessing if the additional cell-type fraction significantly improved the fit, as determined by a likelihood ratio test (LRT) with one degree of freedom ($P < 0.05$). Finally, having obtained an optimal model from UniLIFE and another from the 12 cell-type panel, we then computed the Akaike information criterion $AIC = -2\log[Likelihood] + 2K$ and Bayesian information criterion $BIC = -2\log[Likelihood] + K\log n$, where *K* and *n* are the number of model parameters and datapoints respectively. We computed *AIC* and *BIC* to objectively compare which logistic model is a better predictor of case-control status, as these measured adjust for model complexity.

Robustness of UniLIFE to variations associated with ancestry and genotype

Inference of ancestry DMCs and ancestry effect sizes of UniLIFE CpGs: To infer ancestry-DMCs without the confounding batch effect, we used the DNAm data from Zannas et al. [104], which profiled 393 African Americans and 24 Caucasians. We used *limma* to derive 9,079 ancestry-specific DMCs ($FDR < 0.05$), of which 38 overlapped with UniLIFE probes. To estimate ancestry effect sizes for all UniLIFE CpGs we compared DNAm levels of the UniLIFE CpGs between 3 large cohorts of white Caucasians ($n = 732$, Johansson et al. [105]), Han Chinese ($n = 710$, You et al. [106]) and African Americans ($n = 393$, Zannas et al.).

Simulating impact of ancestry: For all of the UniLIFE CpGs we perturbed their DNAm-values in the Johansson dataset according to a fixed effect determined by the ancestry effect size between Han Chinese and white Caucasians. This simulates the effect of changing the ancestry in the Johansson cohort from whites to Chinese. UniLIFE was then applied to this perturbed data matrix to infer the 19 immune cell fractions. Robustness to ancestry was assessed using the Pearson R-value and RMSE.

mQTLs: We used the list of mQTLs as derived in blood from white Caucasian adults, from the ARIES mQTL database, www.mqtl.org [39] and a separate list of mQTLs as derived in a large Han Chinese cohort [107]. These mQTL lists were overlapped with UniLIFE CpGs.

Simulating impact of genotype: we performed perturbations of the 782 UniLIFE mQTLs from a total of 1880 UniLIFE CpGs present in the Johansson et al. dataset. The procedure was as follows: 1: Convert the beta value matrix to the M value matrix: $M \leftarrow -\log_2(\text{bmiq_simu.m/}$

(1—bmiq_simu.m)) 2: Construct a simulated genotype matrix: `genotypes <- matrix(sample(c(0, 1, 2), size = n_mQTLs * n_samples, replace = TRUE, prob = c(0.5625, 0.375, 0.0625)), nrow = n_mQTLs, ncol = n_samples)`. The three probability values representing the fraction of each genotype were calculated assuming $f_A = 0.75$ and $f_B = 0.25$. 3: Perturbation: `perturbation <- genotypes * mQTL_effect_size`; `M <- M + perturbation`; 4: Convert back to the beta value matrix: `beta_perturbed.m <- 2^M / (1 + 2^M)`; 5: Merge with other CpGs: `beta_perturbed.m <- rbind(beta_perturbed.m, other.m)`. UniLIFE was then applied to this perturbed data matrix to infer the 19 immune cell fractions. Robustness to ancestry was assessed using the Pearson R-value and RMSE.

Implementation

The UniLIFE DNAm reference panel is a matrix of DNAm values defined over 1906 marker CpGs and 19 immune cell-types. Calling this panel “*centUniLIFE.m*”, and given an independent DNAm dataset matrix *X* defined over CpGs (Illumina 450 k, Illumina EPIC1/2 or WGBS) and samples, we estimate fractions for the 19 immune cell-types using the following pseudocode in R-programming language:

```
> library(EpiDISH)
> data(centUniLIFE.m)
> estF19CT.m <- epidish(X, cent = centUniLIFE.m,
  method = "RPC", maxit = 500)$estF
```

where *estF19CT.m* is a matrix containing the estimated immune cell fractions for samples (rows) and each of the 19 immune cell-types. With these estimated fractions it is then possible to (i) identify disease associated differentially methylated cytosines (DMCs) that are not driven by changes in cell-type composition, and (ii) to identify cell-type specific DMCs, as described by us previously [108].

Results

An axis of DNAm variation associated with cell maturity and age

Our overall strategy to build a universal DNAm reference matrix for blood of any age is depicted in Fig. 1a. Briefly, we collected Illumina DNAm data from 7 sorted cord-blood [25, 42, 43, 45] and 12 sorted adult immune cell types [27], and subjected these datasets to a stringent quality control and normalization procedure, removing samples of low purity (Methods, Additional File 2: Fig S1-S2) [29]. This resulted in an integrated DNAm dataset comprising 375,506 CpGs and 220 sorted samples, encompassing 19 immune cell subtypes, 7 from cord-blood and 12 from adult blood (Fig. 1a). PCA-analysis on this merged dataset revealed broad segregation of

samples by cell-type, with PC-1 describing a myeloid vs lymphoid differentiation axis, and with PC-2 describing an axis of variation reflecting cell-maturity (Fig. 1b). Supporting the interpretation of PC-2 as a maturity axis, we note that along this axis, adult naïve B and T-cells are more similar to their cord-blood counterparts than to adult memory B and T-cells. Moreover, for cells from the innate immune system (myeloid and NK-cells), the difference between cord and adult cells was much smaller than that between cord-blood and mature lymphocytes. The PCA plot further revealed that all 19 immune cell-types are separate entities, and that adult naïve T and B-cells are distinguishable from their cord-blood counterparts. Overall, this data indicates that whilst PC-2 is predominantly reflecting a maturity axis, this same axis of variation also discriminates aged naïve/immature lymphocytes from naïve/immature lymphocytes found in cord-blood, suggesting that PC-2 is simultaneously also an axis of ‘aging’.

Construction of a 19 immune cell type DNAm reference matrix from cord and adult blood

Next, we devised a statistical method to select cell-type specific markers, supplementing an empirical Bayes method for detecting differentially methylated cytosines (DMCs) with a novel gap-specificity score and triangulation principle, aimed at maximizing cell-type specificity of the markers (Methods, Additional File 2: Fig S3). Given that the most specific regions for adult cell-types are unmethylated in the given adult cell-type [3], we first aimed to identify at least 100 cell-type specific hypomethylated markers per cell-type. To justify this number, we note that based on our previous experience building a successful DNAm reference panel at the lower resolution of 12 immune cell-types where selecting 50 markers per cell-type performed optimally [29], that increasing cellular resolution from 12 to 19 immune cell-types, would require a concomitant increase in the number of markers per cell-type. For most adult mature cell-types, we could also identify at least 100 cell-type specific hypomethylated markers (Fig. 1c). However, for adult naïve and cord-blood cell subtypes, this was not the case, indicating the need to supplement the marker shortfall with cell-type specific hypermethylated markers (Fig. 1c). To ensure approximately 100 markers for every cell-type (so as to avoid biasing inference to cell-types with many more markers), we also relaxed the gap-specificity score criterion, supplementing the marker shortfall with a criterion based solely on requiring a statistically significant DNAm difference (Methods). This resulted in a total of 1906 marker CpGs across all 19 immune cell subtypes. Finally, a DNAm reference matrix was built by taking the median DNAm value across samples of the same cell-type

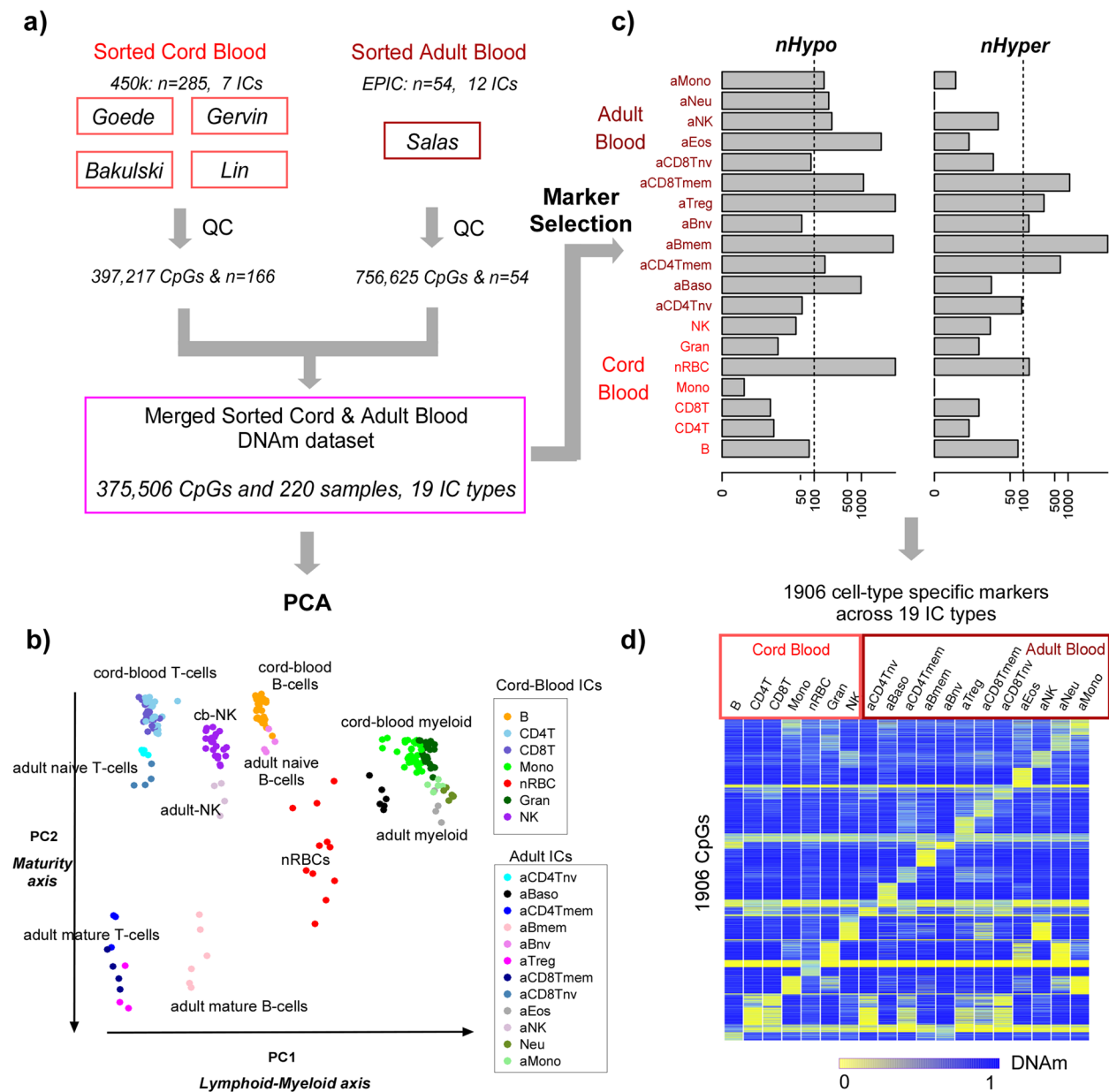


Fig. 1 Construction of the unified cord adult blood DNAm reference matrix. **a** Four independent cord blood Illumina 450 k datasets encompassing 7 immune cell types were processed, underwent separate quality control (QC) analyses which removed samples of low purity. The resulting high quality samples were merged together with an Illumina EPIC DNAm dataset of 12 sorted adult immune cell types, resulting in an integrated DNAm dataset of 220 sorted samples encompassing 19 immune cell-types. **b** PCA scatterplot (PC1 vs PC2) of this merged dataset with samples annotated by one of the 19 immune cell-types. Specific clusters of cells are labeled in the plot to guide the reader. **c** Barplot contrasting the number of cell-type specific markers for each of the 19 immune cell-types, stratified by whether the marker CpG is hypomethylated or hypermethylated in the given cell-type relative to all other cell-types. The x-axis labeling the numbers of markers are displayed on a log-scale. Based on this distribution, a total of 1906 cell-type specific markers were selected. **d** Heatmap representation of the resulting unified DNAm reference matrix for blood encompassing 1906 cell-type specific marker CpGs and 19 immune cell types

(Fig. 1d). We refer to this DNAm reference panel as the UniLIFE (Unified Lifecourse Immune Fraction Estimator) DNAm reference matrix.

Validation in independent sorted and mixed cord blood datasets

To validate the 19 immune cell-type UniLIFE panel, we first assessed its performance in independent cord-blood DNAm datasets. Focusing first on a 450k dataset of 4

sorted cord-blood subtypes, UniLIFE in conjunction with EpiDISH [10] (Methods) predicted high purity for each of these four subsets (Fig. 2a). However, purity was lower for the nRBCs compared to the other 3 cord blood cell-types. Indeed, the nRBCs displayed a non-negligible fraction of cord-blood granulocytes, which could be due to contamination (Fig. 2a). We next applied UniLIFE to sorted cord-blood immune cell samples from IHEC, which were generated with whole-genome bisulfite sequencing. UniLIFE fractions were estimated by restricting to CpGs with 20X coverage (Methods). Of note, the mean number of UniLIFE CpGs covered by at least 20 reads was 1560,

i.e. 82% of all UniLIFE CpGs (Additional File 2: Fig S4a). UniLIFE correctly predicted the cell-type of these sorted samples, with the predictions indicating relatively high purity (Additional File 2: Fig S4b). Importantly, adult blood cell-types were correctly predicted to have zero or near-zero fractions. Upon generating 100 in-silico mixtures, estimated fractions displayed excellent correlation with the true weights (Fig. 2b). Next, we applied our DNAm reference matrix to an independent EPIC DNAm dataset of experimental mixtures of 6 cord blood cell subtypes, where the true mixing fractions are known (Methods), revealing once again an excellent correlation as well

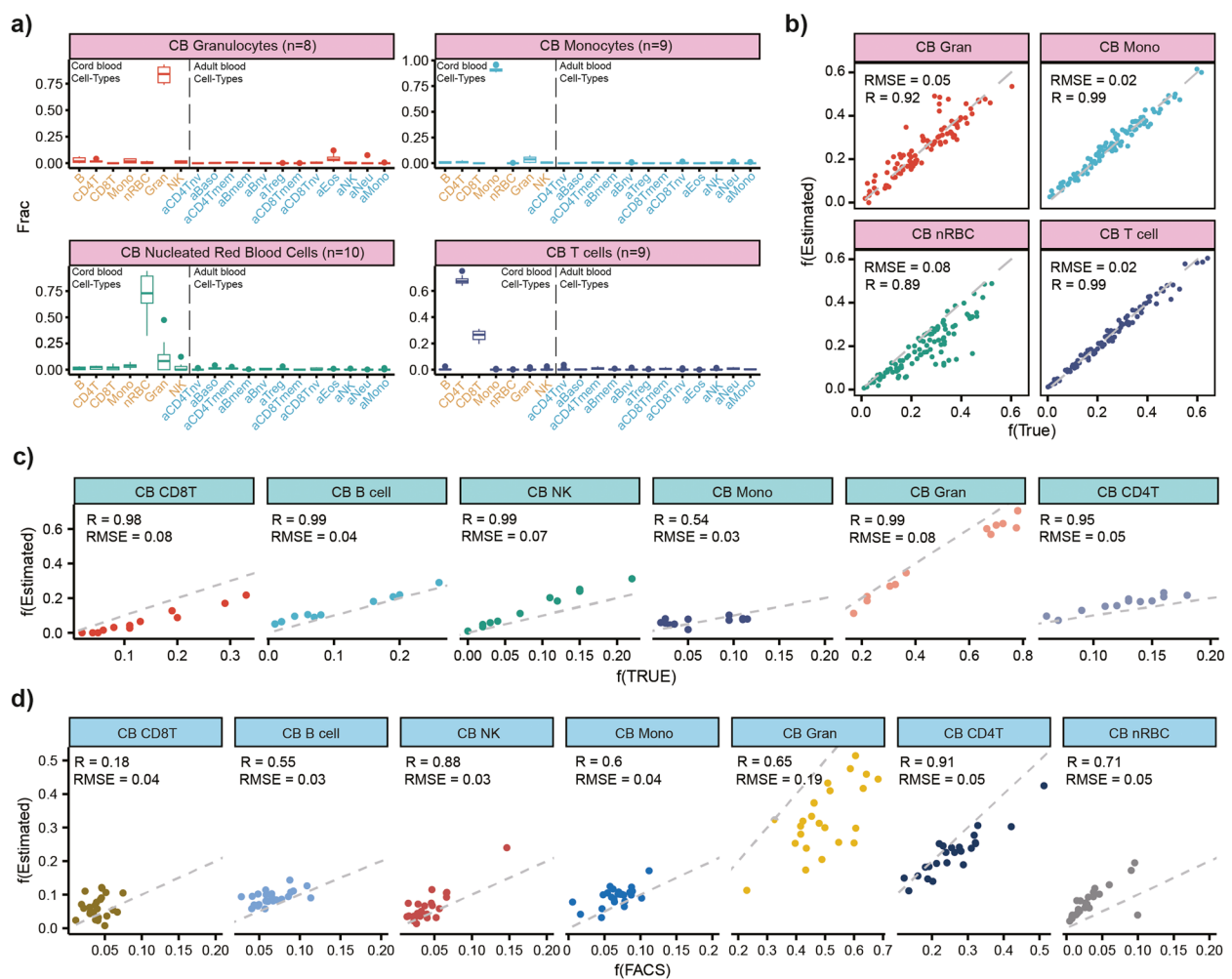


Fig. 2 Validation of UniLIFE in cord blood. **a** Boxplots display the inferred cell type proportions for four sorted cord blood cell types using an independent Illumina 450 k dataset of deGoede. The number of sorted samples corresponding to each cell type is provided. **b** Scatterplots of true mixture weights (x-axis) vs estimated fractions (y-axis) for 100 in-silico mixtures. Each mixture is a linear combination of 4 sorted cord blood samples from a) with weights drawn randomly from a uniform distribution. Pearson R-value and RMSE are given. **c** Scatterplots of true fractions vs estimated fractions for 6 cord blood cell subtypes using an independent EPIC DNAm dataset of 12 umbilical cord blood experimental mixtures where the underlying mixing proportions were known. For each estimated cell type, we display the Pearson R-value and RMSE. **d** Scatterplots of FACS-derived fractions vs estimated fractions for 24 independent cord blood 450 k DNAm profiles with matched FACS cell-counts for 7 cell-types, as shown. Pearson R-value and RMSE are given

as low root mean square errors (RMSE) (Fig. 2c). Only for monocytes was the correlation weaker, which we can attribute to the lower dynamic range of mixing weights used for this particular cell-type. Finally, we applied our DNAm reference matrix to an independent cord-blood 450k DNAm dataset with matched FACS cell counts, including nRBCs. Correlations were strong to very strong (Fig. 2d), except for CD8 + T-cells, which we can attribute to the very low dynamic range of CD8 + T-cell fractions (all FACS counts < 0.08). Indeed, we note that the RMSE for CD8 + T-cells was in fact quite low, suggesting that the lack of correlation is mainly due to the low dynamic range. Importantly, for the nRBC fraction we obtained a reasonably good R and RMSE values. Thus, overall, our DNAm reference matrix can reliably predict fractions in cord blood. To assess robustness of the validation to the number of selected markers per cell-type, we derived another UniLIFE panel using approximately 50 markers per cell-type, resulting in a panel defined over 1020 CpGs and 19 cell-types (Methods). This panel validated as well as the original panel, with no appreciable improvement (Additional File 2: Fig S5), indicating that performance of the 1906 CpG UniLIFE panel is unaffected by a 50% reduction in marker numbers.

Validation in independent sorted and mixed adult blood datasets

Next, we applied a similar strategy to DNAm profiles of adult blood cell-types, experimental mixtures and adult blood with matched experimental cell-counts (Methods). We assembled a large collection of independent Illumina 450 k and whole genome bisulfite sequencing (WGBS) datasets from adult immune cell subtypes, on which UniLIFE correctly predicted the corresponding cell-types (Fig. 3a). Of note, the 7 cord-blood cell subtypes obtained zero or near-zero fractions across all these adult samples, as required (Fig. 3a). To further demonstrate how our DNAm reference matrix can reliably estimate fractions in WGBS data, we generated 100 in-silico mixtures of 10 different WGBS adult blood cell subtypes (restricting again to CpGs with 20X coverage), in each case, revealing an excellent correlation (Fig. 3b). To demonstrate validity on EPIC DNAm data, we compared estimated fractions to experimental mixtures of 10 adult blood cell subtype EPIC DNAm profiles, where the true fractions are known, revealing once again very strong correlations and low RMSEs (Fig. 3c). Finally, in an independent EPIC DNAm dataset of whole blood samples with matched FACS counts, we also observed very strong correlations and low RMSEs despite many fractions displaying a relatively small dynamic range (Fig. 3d). Overall, these results demonstrate that our 19 immune cell-type DNAm reference matrix can reliably estimate fractions in adult blood.

Validation in independent blood datasets of any age

Next, we aimed to demonstrate that UniLIFE can also predict fractions in more complex mixtures composed of cord and adult blood cell-types. This could be relevant for instance in blood of infants or children, where a given leukocyte proportion, say CD4 + T-cells could be modelled as a mix of an immature youthful phenotype, for which those derived from cord-blood could be representative, and an “aged” phenotype encompassing both naïve and mature fractions, as found in adults. We note that this interpretation is well supported by our previous PCA analysis, revealing how PC2 is capturing a maturity axis which is distinct from age, since the adult naïve T-cells cluster more closely to the cord-blood counterparts (Fig. 1b). Thus, we generated in-silico mixtures of cord and adult blood cell-types as a means to further assess the reliability of UniLIFE. Using independent Illumina 450k data of 4 sorted cord blood and 6 sorted adult blood cell-subtypes, we generated such mixtures, revealing that UniLIFE can accurately estimate the known fractions (Fig. 4a). We also generated in-silico WGBS mixtures of 5 cord blood and 10 adult blood cell-subtypes, which once again revealed excellent correlations and low RMSE values (Fig. 4b), demonstrating that UniLIFE can accurately estimate fractions in complex mixtures, independent of underlying technology.

A map of immune-cell fraction variation throughout life reveals non-linear dynamics

To demonstrate practical utility of the 19 immune cell-type DNAm reference matrix, we assembled a large collection of Illumina DNAm datasets encompassing blood tissue from all main age-groups in a human life course, including preterm cord-blood (PreCB), cord-blood (CB), blood from newborns, 6 month old babies, children under the age of 10, adolescents (10–19 year olds), and all age groups up to a maximum of 94 (Methods). Some age-groups contained samples from various cohorts (Additional File 2: Fig S6a), allowing batch effects to be assessed. We observed that batch effects were of a lower magnitude compared to the variation in cell-type fractions (Additional File 2: Fig S6b), allowing the dynamic change of these fractions during a human life course to be delineated. In support of this, we first note how cord-blood cell-type fractions were significantly higher in the PreCB, CB and neonate groups compared to adult samples where the fractions were close to zero (Fig. 5a). The nRBC fraction was also distinctively higher in PreCB and in each of 4 independent CB datasets, as compared to adult cell-types, thus confirming that batch effects are smaller than the expected dynamic changes in cell-type fractions (Additional File 2: Fig S6c). Conversely, adult immune cell-type fractions in PreCB, CB and neonates

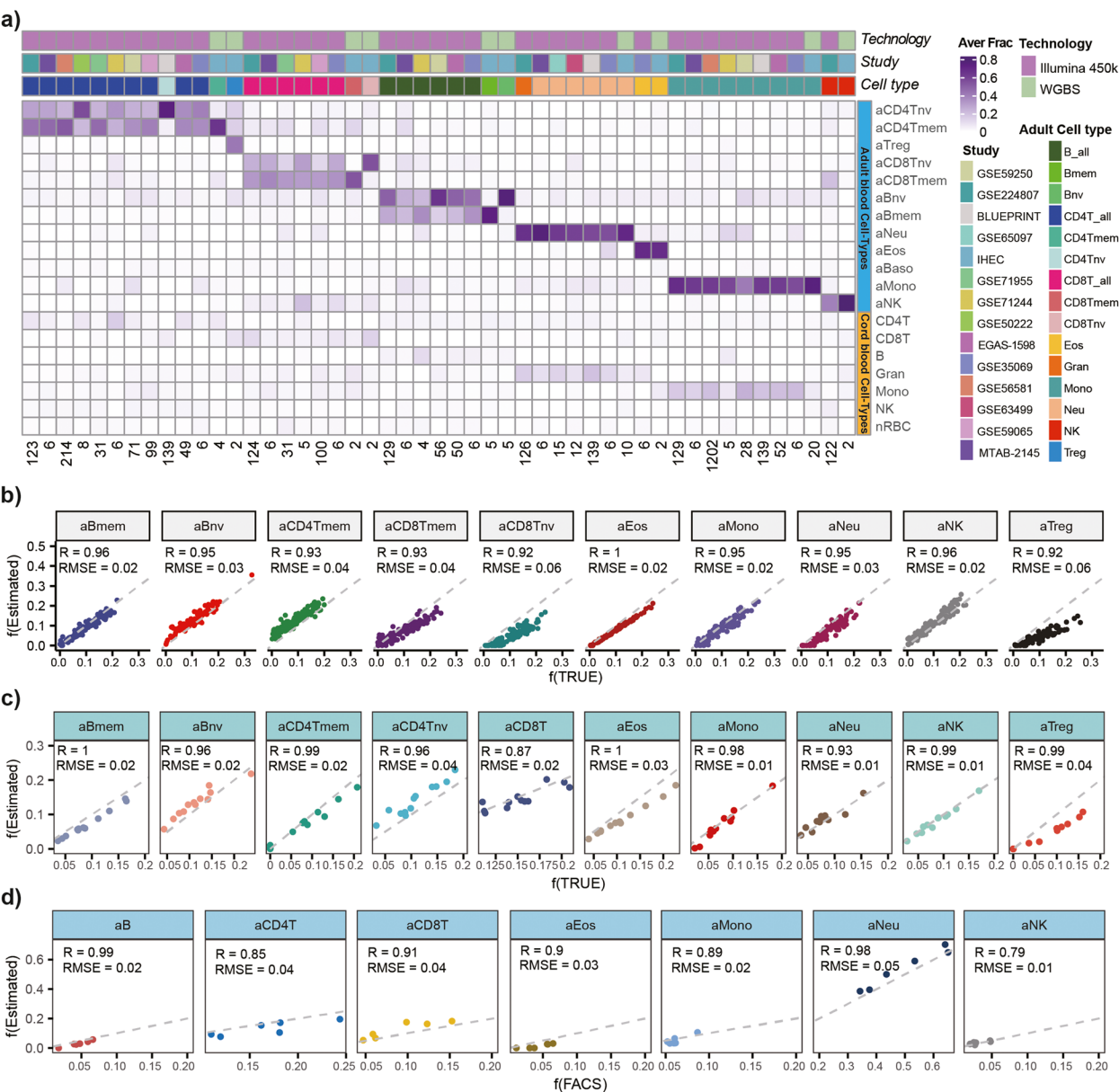


Fig. 3 Validation of UniLIFE in adult blood. **a** Heatmap displays the estimated fractions of sorted samples for each of 12 adult immune-cell subtypes in UniLIFE, as well as the total CD4 + T-cell, total CD8 + T-cell, and total B-cell fractions. The immune-cell type of the sorted sample is indicated by the color bar on top of the heatmap. The technology used to generate the DNAM data of the sorted sample and the study from which the sorted sample derives from are also indicated. The color bar on the right distinguishes between adult (blue) and cord blood (yellow) immune cell types in the UniLIFE reference matrix. The estimated fractions in the heatmap are mean values taken over biological replicates of the cell-sorted samples, with the number of corresponding biological replicate samples indicated at the bottom. **b** Scatterplots of true mixture weights (x-axis) vs estimated fractions (y-axis) for 100 in-silico mixtures of WGBS DNAM profiles from IHEC. Each mixture is a linear combination of 10 adult WGBS immune cell sorted samples with weights drawn randomly from a uniform distribution. Pearson R-value and RMSE are given. **c** Scatterplots of true fractions vs estimated fractions for 10 adult blood cell subtypes using the EPIC DNAM data from 12 experimental mixtures where the underlying mixing proportions were known. For each estimated cell type, we display the Pearson R-value and RMSE. **d** Scatterplots of FACS fractions vs estimated fractions for 6 whole blood EPIC DNAM profiles with matched FACS cell-counts for 7 cell-types, as shown. Pearson R-value and RMSE are given

were very small or zero, subsequently displaying a progressive increase with age, with the exception of naïve lymphocyte fractions, which displayed a non-monotonic pattern, reaching maximum levels during the first 4 years of life (Fig. 5b). The decreased naïve CD4 T/CD8 T fractions with increased age and the corresponding increase

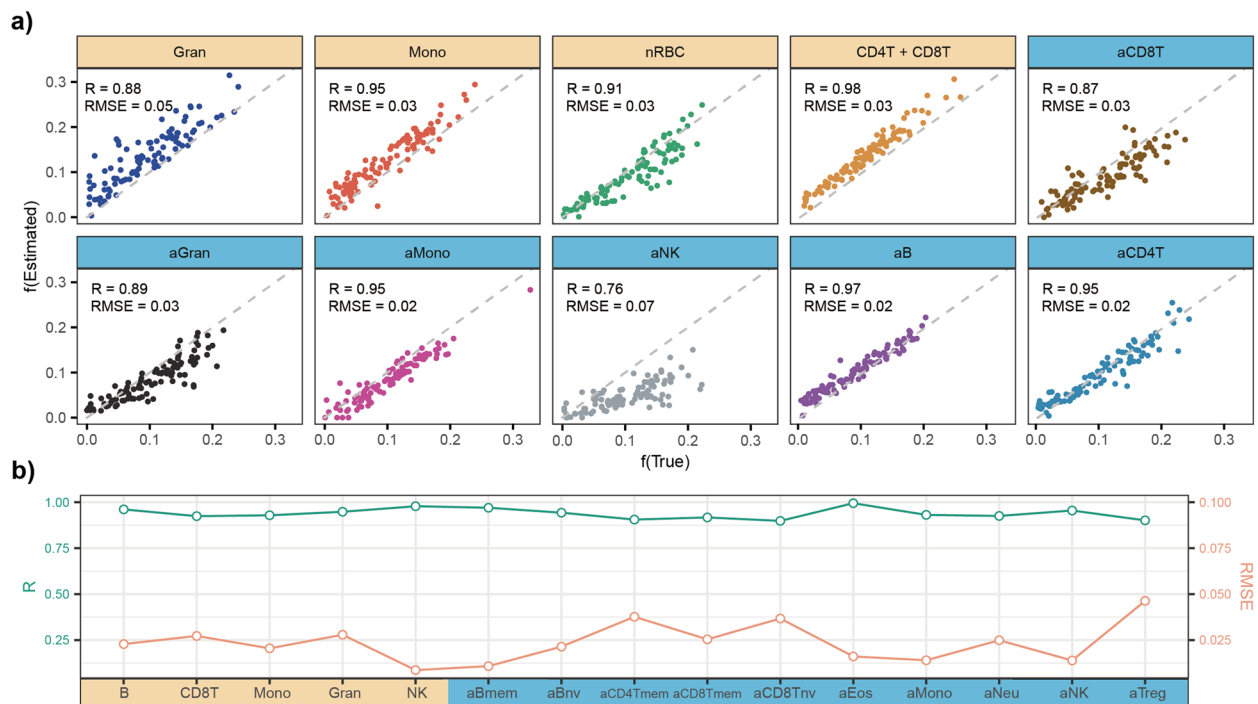


Fig. 4 Validation of UniLIFE in blood of any age. **a** For each of 4 cord blood cell types (orange) and 6 adult blood cell types (blue), scatterplots of true mixture weights (x-axis) vs estimated fractions (y-axis) for 100 in-silico mixtures of independent Illumina 450 k DNAm profiles from Bell and DeGoede datasets. Each mixture is a linear combination of 10 immune cell sorted samples with weights drawn randomly from a uniform distribution. Pearson R-value and RMSE are given. **b** As a), but now for 5 cord blood cell types (orange) and 10 adult blood cell types (blue) with WGBS DNAm profiles from IHEC. This dual-axis line chart illustrates the Pearson R-value (left Y-axis, green line) and RMSE (right Y-axis, orange line) for the true mixture weight and estimated fraction of each cell type across 100 in-silico mixtures. Each mixture is a linear combination of 15 WGBS immune cell sorted samples with weights drawn randomly from a uniform distribution

in memory fractions is well known [109–111]. The rapid decrease in the cord-blood granulocyte fraction within the first 6 months of life is also well known [31], further demonstrating that our inferred patterns are mostly biological and not driven by batch effects. The non-monotonic pattern displayed by the naïve T-cell fractions was robust to the choice of DNAm datasets used, further attesting to the overall robustness of the dynamic patterns to batch effects (Additional File 2: Fig S6 d).

In order to further validate these dynamic DNAm patterns, we compared them to the dynamics of experimentally measured immune-cell type fractions within the first 6 years of life as reported by the Generation R study (GenR) [31] (Methods). First of all, we observed relatively good agreement between our DNAm-based estimated cell-type fractions with those derived by GenR using flow cytometry (Fig. 6a), which is remarkable given that the DNAm and flow cytometric estimates derive from different cohorts. Second, the predicted dynamics from the DNAm-based estimates was well correlated with those derived from flow-cytometry (Fig. 6b-c). For instance, the drop in innate granulocyte and monocyte fractions within the first 6 months of life, as predicted

with our DNAm reference matrix, was also observed using experimental cell counts (Fig. 6b). In the case of naïve lymphocytes, for which our DNAm reference matrix predicted a peak at 6 months, we also observed a corresponding peak with the published experimental counts in the GenR-study (Fig. 6c). Thus, overall, this data supports the view that UniLIFE can accurately chart dynamic changes in cell-type fractions within the first few years of life, an age-group for which we previously did not have a means to quantify such fractions from a DNAm reference matrix. This demonstrates the added-value of our DNAm reference matrix and shows that it is valid for any age-group.

Increased granulocyte to lymphocyte ratio in Down syndrome and preterm sepsis cases

Next, we asked if UniLIFE would be able to make concrete predictions about how immune cell-type proportions change in relation to diseases or conditions that affect infants. First, we considered Down syndrome (DS), since DS patients exhibit widespread immune system dysregulation [112]. Specifically, using complete blood counts, it has been verified that neonates with DS

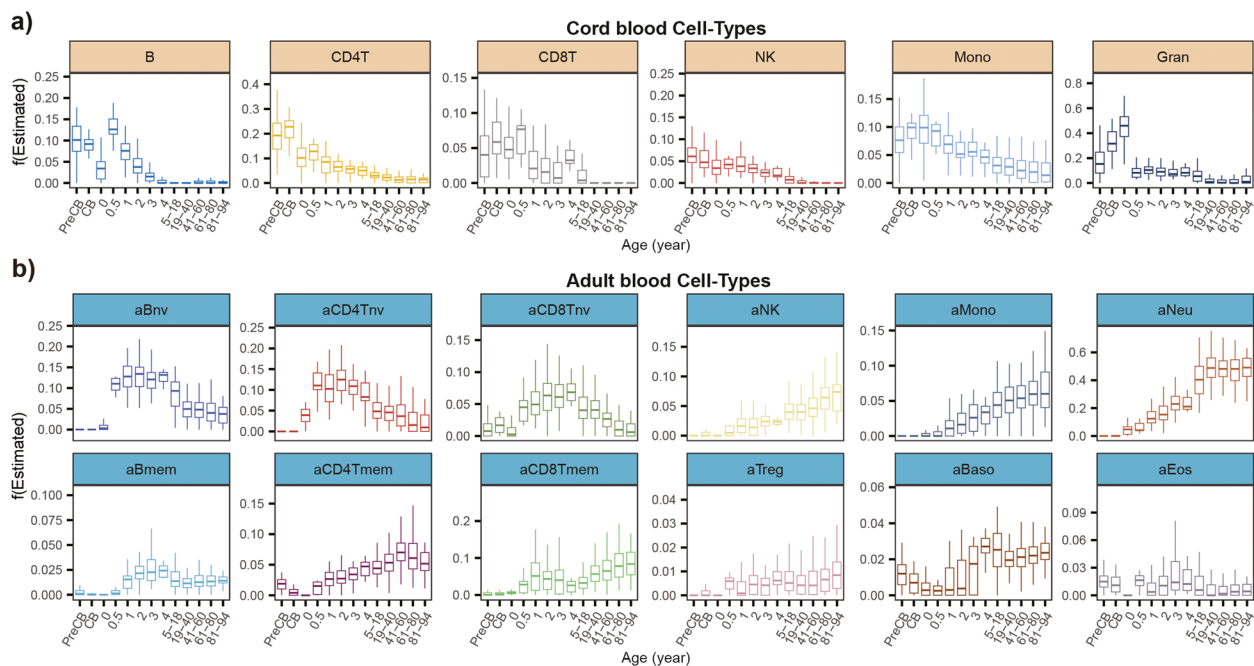


Fig. 5 Immune-cell fraction variation throughout life reveals non-linear dynamics. **a** Boxplots display the dynamic changes of 6 cord blood immune cell type fractions across all life stages. PreCB = cord blood from premature infants, and CB represents cord blood. After the age of 5, life stages are divided into age groups, including 5–18 years, and 20-year intervals thereafter. **b** As a), but for the 12 adult immune cell types

generally exhibit neutrophilia and high nRBC counts including polycythemia [113]. By applying UniLIFE to a HM450k blood DNAm dataset of neonates, encompassing 10 DS cases and 5 non-trisomy controls, we validated the increased cord blood granulocyte and nRBC fractions in DS neonates (Fig. 7a). In contrast, cord blood CD4 T, CD8 T and B-cell fractions were lower in DS cases (Fig. 7a), consistent with flow-cytometric analyses [114]. To see if these changes persist in older DS cases, we estimated immune cell fractions in independent DNAm datasets of toddlers [89] and adults with and without DS [91]. Granulocyte fractions in the older DS cases were no longer different from controls (Fig. 7a), consistent with previous reports [112]. However, the naïve CD4 T, CD8 T and B-cell fractions remained lower in adult DS cases, with the corresponding memory fractions being higher (Fig. 7a). This is consistent with the higher risk of disease and lower life expectancy of DS patients.

Application of the adult 12 immune cell-type reference panel from Luo et al. [29] to the same datasets revealed patterns broadly consistent with those of UniLIFE (Fig. 7b). However, there were also new insights obtained with UniLIFE: for instance, UniLIFE predicts the cord-blood B-cell fraction to be lower in toddlers with DS compared to controls (Fig. 7a), which may indicate a B-cell defect that is therefore already present at birth, consistent with the pattern seen in neonates and with

previous reports [115]. Interestingly, in toddler DS cases, UniLIFE predicts the cord-blood and adult NK-fractions to decrease and increase, respectively, when according to the 12 cell-type panel, the adult NK fraction remains unchanged (Fig. 7a-b). Thus, it would appear that in toddlers with DS the relative proportion of ‘youthful’ NK-cells is lower than in controls, whilst the total NK fraction remains unchanged, which could point towards the cord-blood NK fraction being a surrogate for dysfunctional NK-cells, in line with studies showing that NK-cells in DS cases are functionally impaired (lower cytolytic activity) compared to controls [116].

Next, we applied UniLIFE to an EPIC cord blood DNAm dataset of 23 preterm infants, of which 17 developed sepsis [85]. We observed how the granulocyte and nRBC fractions were significantly increased in sepsis cases compared to controls, whilst correspondingly the T-cell fractions decreased (Additional File 2: Fig S7), consistent with previous reports [117, 118]. We also applied UniLIFE to estimate immune cell fractions in 3 DNAm datasets of toddlers and children, encompassing atopic dermatitis [93], atopic asthma [95] and respiratory allergy [97] cases and age-matched controls (Methods). Although eosinophil and basophil fractions were small, we were able to detect, in each of the 3 conditions, a significant increase of these fractions in cases vs controls (Additional File 2: Fig S8), consistent with the role

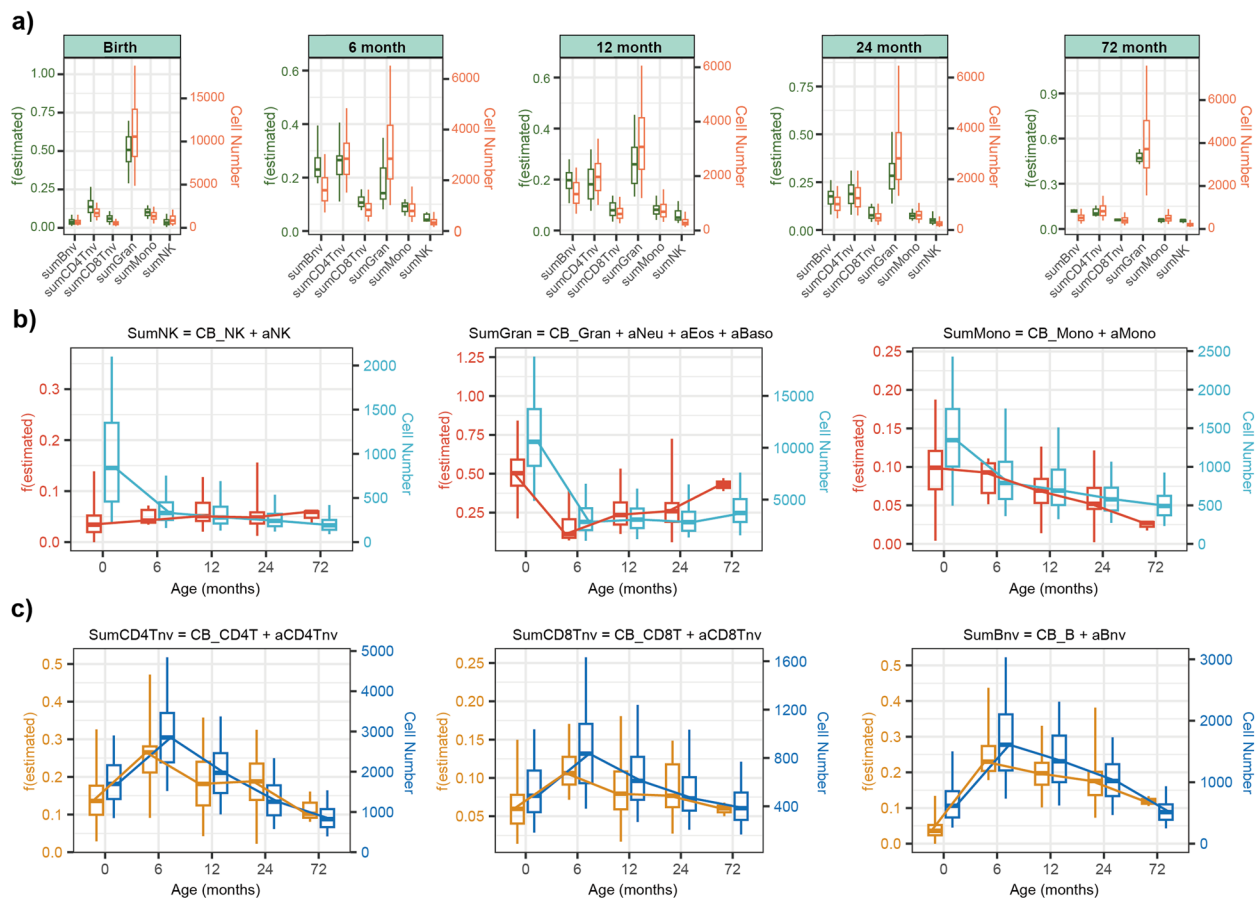


Fig. 6 Comparison between DNAm-based estimates and cell counts reported by GenR. **a** Boxplots compare the DNAm-based estimated fractions of 6 cell types (left y-axis) to experimental cell counts from GenR study (right y-axis) at 5 time points from birth to age 6. The $f(\text{estimated})$ (left y-axis) for each cell type is the sum of the fractions of cord blood type and the corresponding naive adult cell type (for lymphocytes) or the corresponding adult innate type (for myeloid cells), all fractions as estimated from the UniLIFE DNAm reference, i.e. $\text{sumBnv} = \text{CB_B-cell} + \text{aBnv}$, $\text{sumCD4Tnv} = \text{aCD4Tnv} + \text{CB_CD4T}$, $\text{sumCD8Tnv} = \text{aCD8Tnv} + \text{CB_CD8T}$, $\text{sumGran} = \text{CB_Gran} + \text{aNeu} + \text{aBaso} + \text{aEos}$, $\text{sumMono} = \text{CB_Mono} + \text{aMono}$, $\text{sumNK} = \text{CB_NK} + \text{NK}$. Cell Number (right y-axis) refers to the GenR experimental count of cell types. **b** Boxplots compare the consistency of the variation patterns between the DNAm-based estimates and GenR experimental cell counts for three innate cell types measured in the GenR study from ages 0 to 6. The left y-axis, labeled $f(\text{estimated})$, corresponds to the DNAm-based fraction for sumNK, sumGran, and sumMono, while the right y-axis represents the corresponding experimental cell count. **c** As b), but for 3 naive immune cell types

that these immune cell-types play in these diseases [119–122]. Thus, overall, these data demonstrate the ability of UniLIFE to make correct predictions about biological immune-cell fraction variations in infants and children.

UniLIFE infers DNAm changes in MHC-genes that predate the onset of type-1 diabetes

We next showcase how UniLIFE could be used in practice to infer differentially methylated cytosines (DMCs) in a longitudinal prospective study. We focused on an Illumina 450k/EPIC DNAm dataset, which profiled 395 blood samples from infants and children, at multiple timepoints during the natural progression of islet autoimmunity (IA) to type 1 diabetes (T1D) (the DAISY cohort) (Fig. 8a) [99]. This is an excellent case control

study to test UniLIFE on, since prospective T1D cases and controls were matched for all confounders, including age, sex and technical factors (Methods). Using UniLIFE, we estimated the 19 immune cell-type fractions for each sample, which revealed that some fractions (e.g. the adult CD8 + T-cell memory fraction) displayed different dynamic profiles between cases and controls (Fig. 8b). Specifically, whilst the adult CD8 + T-cell memory fraction increased with the timepoint of sampling in controls, it did not display a corresponding increase in the prospective T1D cases. It follows that cell-type heterogeneity, if not adjusted for, could negatively influence the selection of DMCs associated with T1D. To explore this, we ran linear models to find T1D-DMCs, adjusting for the 19 immune cell fractions in addition to other

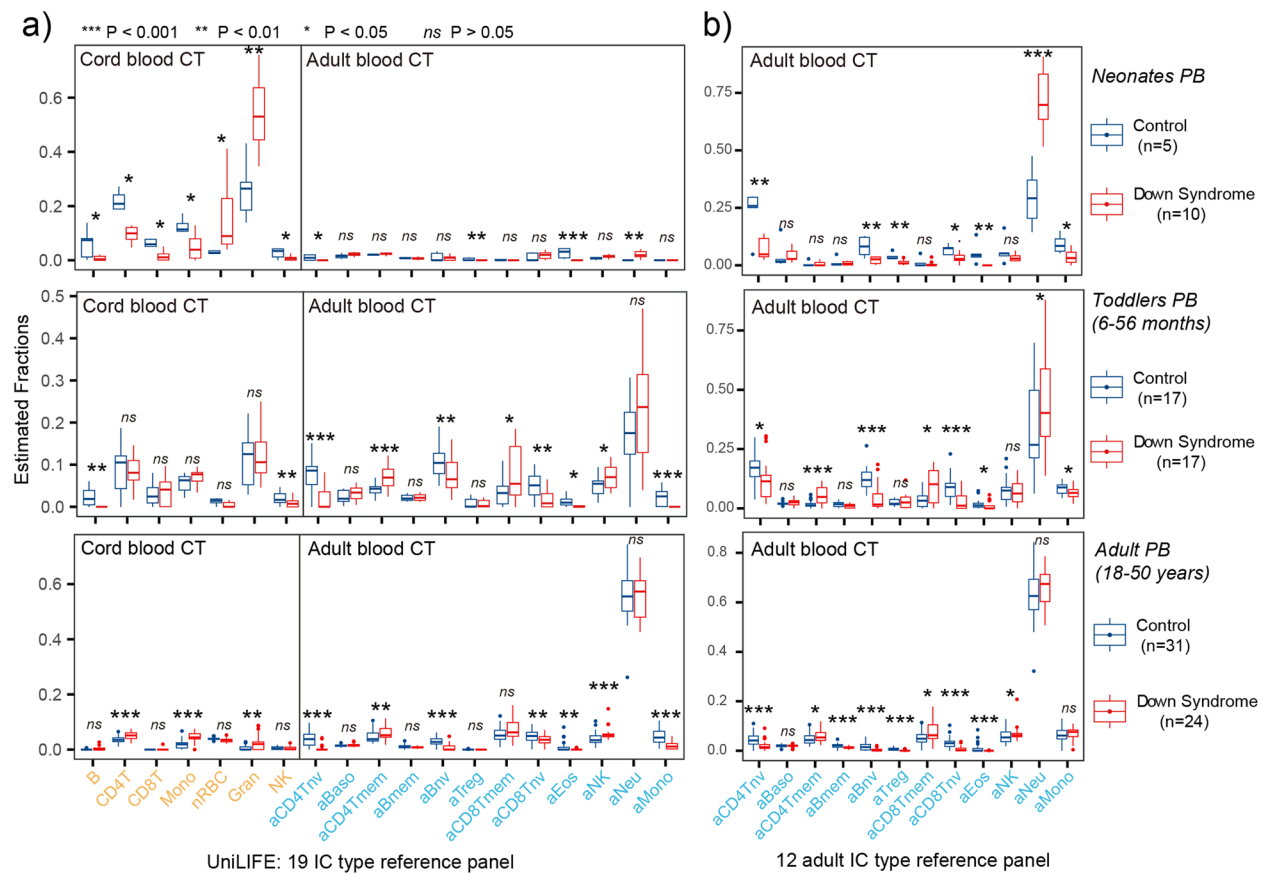


Fig. 7 Validation of biological immune cell fraction variation in Down Syndrome. **a** Boxplot displaying the estimated UniLIFE 19 immune cell type fractions in neonates (top), toddlers (middle) and adults (bottom) with and without Down syndrome. Significance levels are derived from the Wilcoxon rank-sum test, with the symbols *** representing $P < 0.001$, ** representing $P < 0.01$, and * representing $P < 0.05$. **b** As a), but for 12 adult immune cell fractions, as estimated with the 12 adult immune cell-type reference panel from Luo et al.

confounders that included sex and technical factors (chip and position) (Methods). We also adjusted for timepoint, in order to find T1D-DMCs that are consistently differentially methylated irrespective of whether the sample was taken before or after IA. Of note, this was done, because including an interaction term between disease

status and timepoint to find T1D-DMCs where the effect size varies with the time to T1D diagnosis, did not reveal any significant associations, in line with previous findings from the DAISY study [99]. Without an interaction term, we detected 3669 T1D-DMCs at an FDR < 0.05 (Fig. 8c). Using missMethyl [123], we found these to be enriched

(See figure on next page.)

Fig. 8 UniLIFE identifies differential DNAm in T1D pathway before T1D diagnosis. **a** Left barplots compare the number of T1D (top) and control (bottom) samples across 4 distinct timepoints preceding T1D diagnosis in the cases. IA = islet autoimmunity. Right barplots display histograms of the samples according to age. **b** Boxplot displays the estimated fraction of adult CD8 + memory T-cells as a function of timepoint in T1D cases (top) and controls (bottom). P -value is from a linear regression adjusted for age, sex, array-type, chip-position and chip. **c** Volcano plot of T1D-DMCs as identified using a linear model that adjusted for array-type, chip, chip-position, sex, age, timepoint and 19 immune cell fractions as estimated using UniLIFE. **d** Barplots depict the fraction of KEGG pathway genes that were found enriched among the T1D-DMCs derived with the model in c), using fractions estimated with UniLIFE (left), 12 adult immune cell fractions (middle) or no-adjustment (right). The numbers of T1D-DMCs in the latter two cases have been matched to the number obtained using UniLIFE. **e** Violin plots of DNAm profiles of 4 T1D-DMCs mapping to KEGG T1D-pathway members, stratified by prospective disease status, as a function of timepoint. P -value is obtained from the same linear model. **f** Validation of the UniLIFE inferred T1D-DMCs in a monozygotic adult twin study discordant for T1D status (Paul et al.), where samples are sorted monocytes, B-cells or T-cells from approximately 50 twin pairs. The x-axis labels the t-statistics of association with T1D as in the discovery (DAISY) cohort, whilst the y-axis labels the corresponding t-statistics in the validation cohort (Paul et al.). In each panel we count the number of T1D-DMCs that validate at a nominal $P < 0.05$ threshold, or those that display inconsistency. The odds ratio (OR) and P -value from a one-tailed fisher's exact test is given

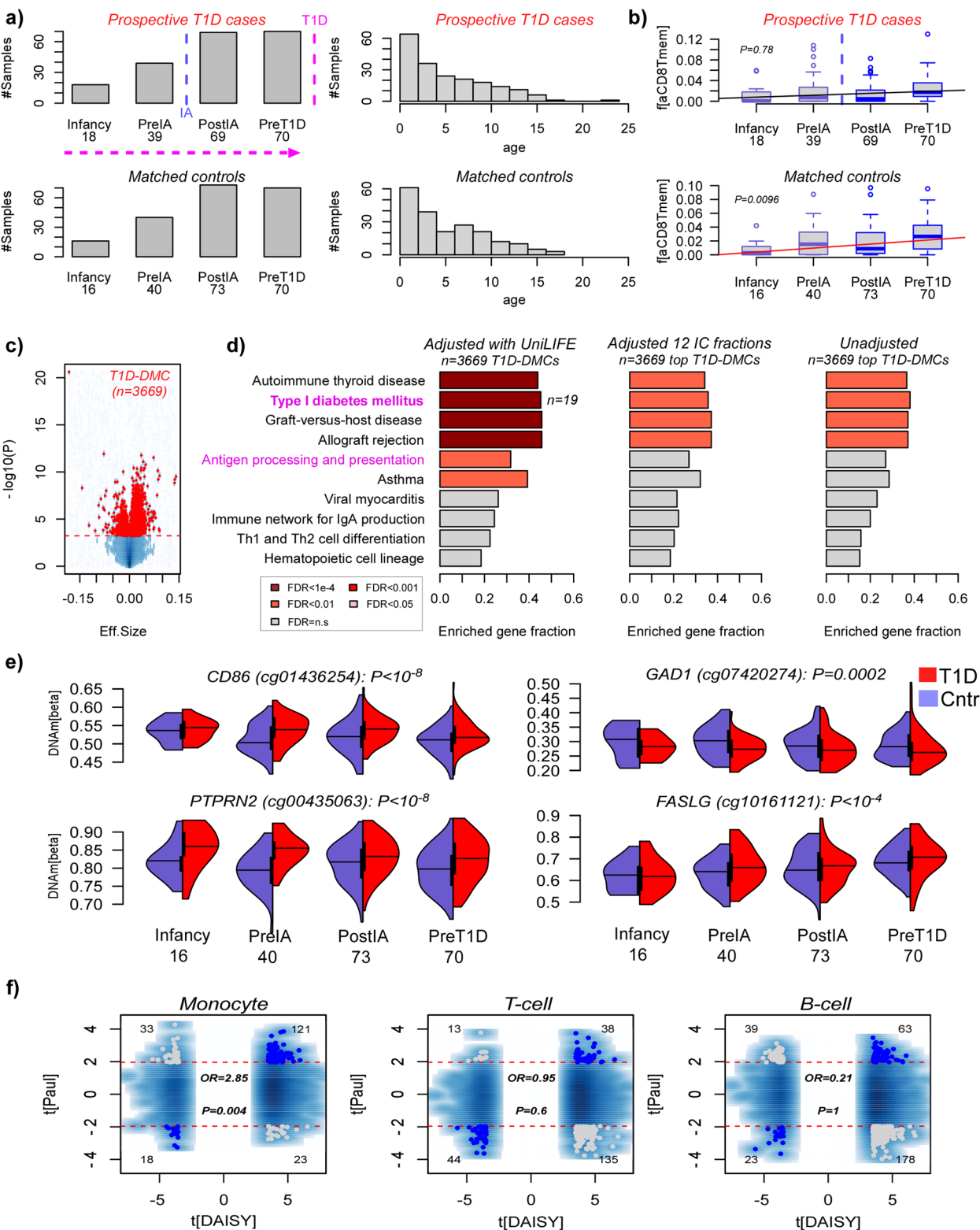


Fig. 8 (See legend on previous page.)

(FDR < 0.0001) for over 40% of the genes in the KEGG type-1 diabetes mellitus pathway (Fig. 8d). Adjusting for only 12 adult immune cell-types, or not adjusting for CTH, did not reveal any significant enrichment, despite the much larger number of T1D-DMCs (Additional File 2: Fig S9), an indication that improper adjustment or no adjustment for CTH, substantially increases the false positive rate. By selecting the top ranked 3669 T1D-DMCs (thus matched for the number obtained with UniLIFE), we also observed enrichment of the KEGG T1D pathway, albeit weaker than when adjusting with UniLIFE (Fig. 8d). Of note, the 19 enriched genes in the T1D KEGG pathway, included members of the MHC-class II (*HLA-DQB1*, *HLA-DQA1/2*, *HLA-DRB1/5*, *HLA-DMB*, *HLA-DOB*), the antigen presenting cell marker *CD86* and genes encoding autoantibodies (*GAD1* and *PTPRN2*). The list also included MHC-class I genes (*HLA-A/B/C/G/F*) and the cytotoxic CD8 + T-cell marker *FASLG* (Fig. 8e) [124]. Unlike the previous DAISY study, we were able to validate our T1D-DMCs (Fig. 8f) in an independent DNAm dataset of sorted monocytes from 50 monozygotic adult twin pairs discordant for T1D [101]. The validation in monocytes, but not in sorted B-cells or T-cells, suggests that the innate immune system may play an important role in the progression of IA to T1D, as recently discussed [124]. Thus, overall, these results support the view that UniLIFE can aid the identification and interpretation of DMCs in longitudinal prospective EWASs performed in infants and children.

UniLIFE improves predictive models of pediatric disease

To further demonstrate the added value of UniLIFE compared to the adult 12 IC-type one, we asked if we can build better predictors of disease using the 19 IC-type fractions compared to only 12. This analysis was performed for 3 separate diseases including atopic asthma (97 controls and 96 cases), respiratory allergy (61 controls and 29 cases) and for Down syndrome (51 cases and 53 controls). In the latter case, in order to have sufficient power, we merged the DS cases and controls from the 3 earlier DS studies encompassing 3 age groups including neonates, toddlers and adults. We then used an 80%–20% training test set strategy, to build elastic net logistic models from the training set with tenfold cross-validation, evaluating the optimal model in the blinded test set. This procedure was repeated in each dataset, a 1000 times with different randomly selected 80%–20% training-test set partitions (Methods). We observed how the model derived from the 19 UniLIFE cell-type fractions obtained higher AUC values than the 12 cell-type panel across 2 of the 3 datasets (allergy + DS), with no substantial difference in the asthma dataset, probably because of a lower number of cases (Additional File 2: Fig S10a). To confirm

that these results are not merely driven by the increased number of variables in the UniLIFE panel, we performed a separate analysis using a forward-selection strategy, comparing final predictive models using the Akaike Information and Bayesian Information Criteria (AIC & BIC), which take into account model complexity (Methods). Confirming the results obtained with the previous training-test set strategy, the UniLIFE model obtained higher log-likelihoods as well as lower AIC/BIC values (Additional File 2: Fig S10b), the latter indicating better predictive models (Methods).

Robustness of the UniLIFE reference panel to missing probes on EPICv2 and MSA arrays

Given that the UniLIFE reference matrix was built from joint 450k-EPICv1 probes, it is important to assess if inferred cell-type fractions are robust to probes that may be missing on the recently released EPICv2 and MSA beadarrays [125, 126]. From the 1906 CpGs in UniLIFE, only 81 are missing or have been flagged in the EPICv2 release. In the case of MSA, 682 UniLIFE probes are not on the MSA, leaving a total of 1,224 CpGs that do have representation. Hence, to simulate performance in the EPICv2 and MSA settings, we re-estimated cell-type fractions removing the 81 and 682 probes, respectively. Across multiple different DNAm datasets, we observed strong agreement between the newly estimated immune cell fractions and those estimated with the full repertoire of UniLIFE CpGs (Additional File 2: Fig S11–S12). Overall, this indicates that inference is robust with the EPICv2 and MSA restricted probe subsets, and hence that UniLIFE would be applicable to new upcoming EPICv2 and MSA datasets.

Generalization of UniLIFE across ancestries and genotypes

Finally, we asked how ancestry and genotype [127, 128] could affect immune cell fraction estimates obtained with UniLIFE. Although the UniLIFE panel itself was built from a mix of different ancestries, the validation sets were predominantly white Caucasian (SI table S1). To assess the potential impact of ancestry, we first inferred ancestry-associated differentially methylated CpGs (ancestry-DMCs) from a DNAm study profiling African Americans and white Caucasians [104]. Out of a total of 9079 ancestry-DMCs, 38 overlapped with UniLIFE probes. For these 38 CpGs we then compared the effect sizes (absolute DNAm difference) associated with ancestry to the effect sizes associated with cell-type (as given by the UniLIFE panel) (Additional File 2: Fig S13a). Whilst the ancestry effect sizes were effectively all less than 10% (i.e. < 0.1 on the beta-value scale), cell-type specific effect sizes were typically > 50%. To see if this generalizes, we estimated ancestry effect sizes for all 1906 UniLIFE CpGs

by comparing DNAm-levels between 3 main ancestries (White Caucasians, Han Chinese and African Americans) (Methods). As before, effect sizes associated with cell-type were substantially larger (about 10 times larger) than those associated with ancestry (Additional File 2: Fig S13b). Based on this, ancestry should have a minimal impact on the inference of immune cell fractions with UniLIFE. To confirm this, we simulated the impact of ancestry by perturbing the DNAm profiles of the 1906 UniLIFE probes in a white Caucasian DNAm dataset ($n = 732$, Johansson et al. [105]), perturbing them according to their ancestry effect sizes as estimated by comparing Han-Chinese to white Caucasians (Methods). Thus, the perturbed dataset over the UniLIFE probes can be viewed as a simulated Han Chinese cohort. Next, we asked how similar the cell-type fractions inferred in this simulated Han-Chinese cohort are to those of the original white Caucasian dataset. We observed a strong agreement (Additional File 2: Fig S13c), with only some minor differences for the rarer cell-types, suggesting that, overall, UniLIFE should be applicable across ancestries.

Next, we performed a similar analysis for methylation quantitative trait loci (mQTLs). Around 30% of human CpGs constitute mQTLs [107, 129]. Thus, not unsurprisingly, we found a substantial overlap of UniLIFE CpGs with mQTLs: for instance, 796 UniLIFE CpGs constitute mQTLs in a large Han Chinese population [107]. Nevertheless, as with overlapping ancestry-DMCs, the mQTL effect size of overlapping UniLIFE-mQTL CpGs was much smaller ($< 10\%$) than the effect sizes associated with cell-type ($> 50\%$) (Additional File 2: Fig S14a). Performing a similar analysis as for ancestry, we next simulated the impact of variable genotypes on the inference of cell-type fractions (Methods). UniLIFE's immune cell fractions were highly correlated between the original and genotype-perturbed data (Additional File 2: Fig S14b), supporting the view that inference of these fractions with UniLIFE is robust to variations in genotype.

Discussion

Here we have advanced the field, by contributing a DNAm reference panel that works for blood tissue of any age. Although it was built using only sorted samples from cord and adult blood cell-types, these can be viewed as extremes in a continuum that includes blood from infants, children and adolescents. A key insight supporting this interpretation was derived by studying the principal components of variation of the merged DNAm matrix: this PCA revealed that adult naïve lymphocytes are more similar to their cord-blood counterparts than they are to adult (age-matched) mature lymphocytes, indicating that we have successfully identified a maturity axis of DNAm variation. Importantly, this same axis

of variation could also discriminate, albeit with a lower effect size, cord blood from adult innate immune cell-types, including granulocytes, monocytes and NK-cells. Thus, this same maturity axis of DNAm variation is simultaneously informative of age. As such, UniLIFE is able to chart the dynamic landscape of major immune cell-type proportions during the whole human life course.

We have extensively validated UniLIFE on (i) independent sorted cord and adult blood cell-types, (ii) on unsorted samples with matched experimental cell-counts and (iii) in-silico mixtures, encompassing two independent technologies (WGBS and Illumina EPIC/450 k). Most importantly, UniLIFE makes predictions of dynamic immune cell type compositional changes within the first months and years of life that recapitulate those derived with experimental cell-counts in the GenR study. For instance, that naïve lymphocyte fractions reach a maximum at 6 months, remaining stable for the first 2 years of life, subsequently declining with age, as the corresponding memory fractions increase, is entirely consistent with those previously reported in a large study featuring repeatedly assessed cell counts [31]. Further validating UniLIFE in infants and children, we have shown how we can correctly predict well-established dynamic cell-type compositional changes in a wide range of diseases and conditions, including Down syndrome, atopic dermatitis, asthma and respiratory allergy. For instance, we confirmed an increased granulocyte and nRBC fraction in neonates with DS compared to non-trisomy controls, but that this increase is no longer observed in children and adult DS cases. In contrast, naïve T and B-cell fractions, which are distinctively reduced in DS neonates, remain lower in children and adult DS cases compared to age-matched controls, which may underpin their lower life expectancy. Thus, the fact that UniLIFE can correctly predict these dynamic patterns for immune cell fractions during the first years of life, including adolescence, not only demonstrates the reliability of the estimated fractions, but also demonstrates its practical utility beyond what could have been achieved with a cord or adult-blood only DNAm reference matrix.

We have also showcased how UniLIFE can be applied to an EWAS performed in infants and children, to identify T1D-associated DMCs that predate the diagnosis of T1D. Specifically, our analysis revealed how specific immune cell-fractions (e.g. adult CD8 + memory cells) display different dynamics in prospective T1D cases compared to age-matched controls, indicating that proper adjustment for CTH could be critical. In line with this, our supervised analysis adjusting for 19 immune cell type fractions led to a stronger statistical enrichment for genes implicated in the T1D-pathway,

despite a smaller number of T1D-DMCs, suggesting that improper adjustment for CTH using lower cellular resolution, or not performing any adjustment, leads to a larger number of false positives, consistent with previous reports on how CTH can inflate the false positive rate [7, 8]. The fact that using UniLIFE we could prospectively identify T1D-DMCs mapping to 19 genes in the T1D-pathway is highly significant, specially given that this was not reported by the original DAISY study [99]. Moreover, unlike the original DAISY study, we were able to validate the pattern of differential methylation in association with T1D in an independent adult monozygotic twin study discordant for T1D status [101]. Interestingly, the validation was seen in sorted monocytes, but not in B or T-cells, which could reflect the prospective nature of our T1D-DMCs, as it has been recently argued that innate immune cells play a more important role in the stages before T1D diagnosis [124]. Although the monozygotic twin study predominantly comprises adults with diagnosed T1D (i.e. samples are postdiagnosis), it is plausible that DNAm changes in monocytes that predispose to T1D would still be present after diagnosis. Although this will require further investigation, our data clearly indicates how UniLIFE can lead to novel biological results and hypotheses.

It is important to end with a number of limitations. Although we have extensively validated our DNAm reference matrix, it would be desirable to have validated it in more independent sorted nRBC samples. We note that the sorted nRBC samples used in the construction of the DNAm reference matrix were more dispersed in the PCA plot, possibly reflecting subtle contamination by diverse white blood cell-types [43, 130]. Indeed, nRBCs are known to interact with white blood cells, forming doublets that can result in cross-contamination when using conventional isolation techniques such as FACS or MACS [43, 130]. Thus, future improved iterations of DNAm reference matrices like UniLIFE would benefit from larger collections of sorted cord blood cell-types of higher purity, so as to increase the reliability and accuracy of the estimated fractions.

It is also worth pointing out that our UNILIFE DNAm reference matrix was built using joint 450 k-EPICv1 probes. As such, further improvements could be envisaged if all sorted cord and adult blood data could have been generated with the same and more comprehensive EPIC platform. Such future attempts are very likely to use the recently released EPICv2 platform [125]. With regard to this, we note that although approximately 17% of EPICv1 probes are not present on EPICv2, less than 5% (81 probes) of the 1906 CpGs in our UNILIFE reference panel are missing or have been flagged as unreliable

from the EPICv2 array. As shown by us previously [131], and demonstrated again here, the inference of cell-type fractions from a DNAm reference matrix is very robust to missing probes or putative errors, tolerating up to 20–30% of missing data/errors without compromising reliability. This is because of the multivariate nature of the inference procedure, which is tantamount to a ‘voting or wisdom of crowds algorithm’, the implication being that as long as the majority of the entries in the reference matrix are present and reasonably accurate, the algorithm will converge to the correct solution. Consequently, and as explicitly demonstrated here, UniLIFE is applicable to EPICv2. Similar arguments apply to the recently released MSA array [126], which is substantially sparser and therefore contains far fewer but still a significant number of UniLIFE probes ($n = 1,224$). Indeed, as shown here, estimated immune cell fractions were very robust when restricting to these 1224 joint 450 k-EPICv1-MSA probes, supporting the view that UniLIFE is also applicable to upcoming MSA data. Finally, we have shown that UniLIFE is also applicable to other technologies such as WGBS. Although read-depth is a critical factor, it is worth stressing again that UniLIFE was built with around 100 CpGs per cell-type, rendering it very robust to missing data. Supporting this, the successful validation on in-silico mixtures of sorted IHEC WGBS samples, which was performed without restrictions on read coverage, points towards accurate inference even in the presence of unreliable measurements, yet we acknowledge that for WGBS generated at lower coverages (e.g. 5–10X), inference could be compromised.

The UniLIFE panel also demonstrated substantial robustness to variations in ancestry and genotype. Whilst the panel itself was derived from a mix of different ancestries, most of the validation datasets used were from White Caucasians. Nevertheless, we were able to simulate the effect of ancestry and genotype, demonstrating that variation induced by these factors only has a relatively minor impact on the cell-type fraction estimates. Once again, this can be understood by the fact that UniLIFE CpGs overlapping with ancestry and genotype DMCs display effect sizes in relation to ancestry and genotype that are on average around 10 times smaller than those they display between immune cell-types. Nevertheless, for rarer cell-types where the margin of error in cell-type fraction estimation is much smaller, ancestry could have a non-negligible impact. Hence, it might be important for future efforts to build reference panels from a larger number of sorted samples drawn from donors representing all major ancestries.

A final limitation worth stressing is that UniLIFE was built from 19 sorted cell-types selected with specific surface markers, which may not be 100% specific. Moreover,

there are substantially more than 19 immune cell types in the human immune system (by some estimates several hundreds) [132], with the number of functionally distinct immune cell-types also varying depending on the age of the subject. As such, the estimated cell-type fractions from UniLIFE cannot fully adjust for all underlying cell-type heterogeneity. Looking forward, it will be important to characterize the most relevant immune-cell subtypes for aging and disease, guided by single-cell studies, and to build correspondingly higher resolution DNAm reference panels.

Conclusions

In summary, this study contributes a novel epigenetic software tool to enable unified estimation of 19 immune-cell type fractions across a human life-course. It is applicable to blood of any age, including cord-blood from preterm infants and newborns, as well as blood from infants, children, adolescents and adults. We envisage that this resource will enhance and aid interpretation of cross-cohort and longitudinal EWAS that aim to better model developmental/time-varying epigenetic signals in immune cell-types, by better dissecting the confounding effect of cell-type heterogeneity. The estimated immune cell fractions from UniLIFE may also serve as predictive or prognostic biomarkers. The UniLIFE DNAm reference matrix has been incorporated into our EpiDISH R-package, which is freely available from <https://www.bioconductor.org/packages/devel/bioc/html/EpiDISH.html> or <https://github.com/sjczheng/EpiDISH>.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-025-01489-7>.

Additional file 1. An excel table (ADF1.xls) including Table S1 and its caption.

Additional file 2. A pdf document (ADF2.pdf) containing all Supplementary Figures and their legends.

Availability and requirements

Project name: UniLIFE within EpiDISH.

Project home page: <https://bioconductor.org/packages/release/bioc/html/EpiDISH.html>.

Operating system(s): Platform independent.

Programming language: R(> 4.0.0).

License: GNU GPL v3.

Any restrictions to use by non-academics: none.

Authors' contributions

XG performed the statistical and bioinformatic analyses. AET, MS and AN assisted with bioinformatic analyses and literature curation. SCZ updates and maintains EpiDISH. BTH, AET and CAMC conceived the study and wrote the article. All authors read and approved the final manuscript.

Funding

AET was supported by NSFC (National Science Foundation of China) grants, grant numbers 32170652 and 32370699, and an RFIS grant W2431024. CAMC is supported by the HorizonEurope Research and Innovation Program (HappyMums, grant agreement No 101057390; FAMILY, grant agreement No 101057529; STAGE, grant agreement No.101137146) and the European Research Council (TEMPO; grant agreement No 101039672). This research was conducted while CAMC was a Hevolution/AFAR New Investigator Awardee in Aging Biology and Geroscience Research.

Data availability

The following DNAm datasets analyzed here are publicly available from the NCBI GEO website under the accession numbers GSE68456 (de Goede_1), GSE82084 (de Goede_2), GSE180970 (Salas), GSE127824 (Jones), GSE149572 (Hernandez), GSE269983 (Puvvula), GSE97628 (Holland), GSE188949 (Bell_1), GSE36064 (Barwick), GSE224573 (Quinn), GSE104812 (Wang), GSE224807 (Bell_2), GSE112618, GSE167998, GSE56581, GSE65097, GSE50222, GSE63499, GSE71955, GSE71244, GSE59250, GSE59065, GSE35069, GSE155952 (Lorentepozo), GSE107211 (Henneman), GSE174555 (Naumova), GSE52588 (Bacalini), GSE152084 (Chen), GSE40576 (Yang), GSE110128 (Langie), GSE40279 (HNM), GSE210254 (HPT-450k), GSE72680 (Zannas), GSE87571 (Johansson), GSE142512 (DAISY). The sorted dataset of Zilbauer et al. can be accessed at ArrayExpress under accession number E-MTAB-2145. BLUEPRINT data can be accessed at EUROPEAN GENOME-PHENOME ARCHIVE (EGA) under accession number EGAD00010000850. And the sorted dataset of Paul et al. is also available from the EGA under accession number EGAS00001001598. International Human Epigenome Consortium data can be accessed at the IHEC data portal (<https://epigenomesportal.ca/ihec/>).

Declarations

Ethics approval and consent to participate

We only analyzed publicly available data.

Consent for publication

Not applicable.

Competing interests

AET is a consultant for TruDiagnostics Inc. SCZ is an employer of Pfizer.

Received: 9 January 2025 Accepted: 16 May 2025

Published online: 27 May 2025

References

- Lappalainen T, Greal JM. Associating cellular epigenetic models with human phenotypes. *Nat Rev Genet*. 2017;18:441–51.
- Moss J, Magenheimer J, Neiman D, Zemmour H, Loyfer N, Korach A, Samet Y, Maoz M, Druid H, Arner P, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. *Nat Commun*. 2018;9:5068.
- Loyfer N, Magenheimer J, Peretz A, Cann G, Bredno J, Klochendler A, Fox-Fisher I, Shabi-Porat S, Hecht M, Pelet T, et al. A DNA methylation atlas of normal human cell types. *Nature*. 2023;613:355–64.
- Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR, et al. The NIH roadmap epigenomics mapping consortium. *Nat Biotechnol*. 2010;28:1045–8.
- Stunnenberg HG, International Human Epigenome C, Hirst M: The international human epigenome consortium: a blueprint for scientific collaboration and discovery. *Cell*. 2016;167:1145–1149.
- Qi L, Teschendorff AE. Cell-type heterogeneity: why we should adjust for it in epigenome and biomarker studies. *Clin Epigenetics*. 2022;14:31.
- Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol*. 2014;15:R31.
- Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M, et al. Epigenome-wide

- association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol.* 2013;31:142–7.
9. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13: 86.
 10. Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC Bioinformatics.* 2017;18:105.
 11. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nat Rev Genet.* 2018;19:129–47.
 12. Zheng SC, Beck S, Jaffe AE, Koestler DC, Hansen KD, Houseman AE, Irizarry RA, Teschendorff AE. Correcting for cell-type heterogeneity in epigenome-wide association studies: revisiting previous analyses. *Nat Methods.* 2017;14:216–7.
 13. Li Z, Wu Z, Jin P, Wu H. Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics.* 2019;35:3898–905.
 14. Fraser A, Macdonald-Wallis C, Tilling K, Boyd A, Golding J, Davey Smith G, Henderson J, Macleod J, Molloy L, Ness A, et al. Cohort profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort. *Int J Epidemiol.* 2013;42:97–110.
 15. Kirchenko P, Lahti M, Tuovinen S, Savolainen K, Lahti J, Binder EB, Reynolds RM, Entringer S, Buss C, Wadhwa PD, et al. Cohort profile: prediction and prevention of preeclampsia and intrauterine growth restriction (PREDO) study. *Int J Epidemiol.* 2017;46:1380–1381g.
 16. Guxens M, Ballester F, Espada M, Fernandez MF, Grimalt JO, Ibarluzea J, Olea N, Rebagliato M, Tardon A, Torrent M, et al. Cohort profile: the INMA–Infancia y Medio Ambiente–(Environment and Childhood) project. *Int J Epidemiol.* 2012;41:930–40.
 17. Neumann A, Walton E, Alemany S, Cecil C, Gonzalez JR, Jima DD, Lahti J, Tuominen ST, Barker ED, Binder E, et al. Association between DNA methylation and ADHD symptoms from birth to school age: a prospective meta-analysis. *Transl Psychiatry.* 2020;10:398.
 18. Relton CL, Gaunt T, McArdle W, Ho K, Duggirala A, Shihab H, Woodward G, Lyttleton O, Evans DM, Reik W, et al. Data resource profile: Accessible Resource for Integrated Epigenomic Studies (ARIES). *Int J Epidemiol.* 2015;44:1181–90.
 19. Kooijman MN, Kruithof CJ, van Duijn CM, Duijts L, Franco OH, van IMH, de Jongste JC, Klaver CC, van der Lugt A, Mackenbach JP, et al. The generation R study: design and cohort update 2017. *Eur J Epidemiol.* 2016;31:1243–1264.
 20. Felix JF, Joubert BR, Baccarelli AA, Sharp GC, Almqvist C, Annesi-Maesano I, Arshad H, Baiz N, Bakermans-Kranenburg MJ, Bakulski KM, et al. Cohort profile: Pregnancy And Childhood Epigenetics (PACE) Consortium. *Int J Epidemiol.* 2018;47:22–23u.
 21. Cecil CAM, Neumann A, Walton E. Epigenetics applied to child and adolescent mental health: progress, challenges and opportunities. *JCPP Adv.* 2023;3:e12133.
 22. Rakyant VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. *Nat Rev Genet.* 2011;12:529–41.
 23. van den Broek T, Borghans JAM, van Wijk F. The full spectrum of human naive T cells. *Nat Rev Immunol.* 2018;18:363–73.
 24. Reinus LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS ONE.* 2012;7: e41361.
 25. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, Duijts L, Moll HA, Kelsey KT, Kobor MS, et al. Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. *Clin Epigenetics.* 2019;11:125.
 26. Salas LA, Koestler DC, Butler RA, Hansen HM, Wiencke JK, Kelsey KT, Christensen BC. An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina Human-MethylationEPIC BeadArray. *Genome Biol.* 2018;19:64.
 27. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC. Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. *Nat Commun.* 2022;13:761.
 28. Gervin K, Page CM, Aass HC, Jansen MA, Fjeldstad HE, Andreassen BK, Duijts L, van Meurs JB, van Zelm MC, Jaddoe VW, et al. Cell type specific DNA methylation in cord blood: a 450K-reference data set and cell count-based validation of estimated cell type composition. *Epigenetics.* 2016;11:690–8.
 29. Luo Q, Dwaraka VB, Chen Q, Tong H, Zhu T, Seale K, Raffaele JM, Zheng SC, Mendez TL, Chen Y, et al. A meta-analysis of immune-cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes. *Genome Med.* 2023;15:59.
 30. Fernando KK, Craig JM, Dawson SL. Relationships between the maternal prenatal diet and epigenetic state in infants: a systematic review of human studies. *J Dev Orig Health Dis.* 2023;14:540–55.
 31. van den Heuvel D, Jansen MAE, Nasserinejad K, Dik WA, van Lochem EG, Bakker-Jonges LE, Bouallouch-Charif H, Jaddoe VWV, Hooijkaas H, van Dongen JJM, et al. Effects of nongenetic factors on immune cell dynamics in early childhood: the Generation R Study. *J Allergy Clin Immunol.* 2017;139(1923–1934): e1917.
 32. Mulder RH, Neumann A, Cecil CAM, Walton E, Houtepen LC, Simpkin AJ, Rijlaarsdam J, Heijmans BT, Gaunt TR, Felix JF, et al. Epigenome-wide change and variation in DNA methylation in childhood: trajectories from birth to late adolescence. *Hum Mol Genet.* 2021;30:119–34.
 33. Mulder RH, Neumann A, Felix JF, Suderman M, Cecil CAM. Characterising developmental dynamics of adult epigenetic clock sites. *EBioMedicine.* 2024;109: 105425.
 34. Neumann A, Sammalhti S, Cosin-Tomas M, Reese SE, Suderman M, Alemany S, Almqvist C, Andrusaityte S, Arshad SH, Bakermans-Kranenburg MJ, et al. Epigenetic timing effects on child developmental outcomes: a longitudinal meta-regression of findings from the Pregnancy And Childhood Epigenetics Consortium. *Genome Med.* 2025;17:39.
 35. Silk T, Dipnall L, Wong YT, Craig JM. Epigenetics and ADHD. *Curr Top Behav Neurosci.* 2022;57:269–89.
 36. Marioni RE, Suderman M, Chen BH, Horvath S, Bandinelli S, Morris T, Beck S, Ferrucci L, Pedersen NL, Relton CL, et al. Tracking the epigenetic clock across the human life course: a meta-analysis of longitudinal cohort data. *J Gerontol A Biol Sci Med Sci.* 2019;74:57–61.
 37. Arathimos R, Granell R, Henderson J, Relton CL, Tilling K. Sex discordance in asthma and wheeze prevalence in two longitudinal cohorts. *PLoS ONE.* 2017;12: e0176293.
 38. Kuo PL, Moore AZ, Tanaka T, Belsky DW, Lu ATH, Horvath S, Bandinelli S, Ferrucci L. Longitudinal changes in epigenetic clocks predict survival in the InCHIANTI cohort. *medRxiv.* 2024. <https://doi.org/10.1101/2024.09.13.24313620>.
 39. Gaunt TR, Shihab HA, Hemani G, Min JL, Woodward G, Lyttleton O, Zheng J, Duggirala A, McArdle WL, Ho K, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome Biol.* 2016;17:61.
 40. Martino D, Kresoje N, Amenyogbe N, Ben-Othman R, Cai B, Lo M, Idoko O, Odumade OA, Falsafi R, Blimkie TM, et al. DNA Methylation signatures underpinning blood neutrophil to lymphocyte ratio during first week of human life. *Nat Commun.* 2024;15:8167.
 41. Martino D, Schultz N, Kaur R, van Haren SD, Kresoje N, Hoch A, Diray-Arce J, Su JL, Levy O, Pichichero M, in association with the IC: Respiratory infection- and asthma-prone, low vaccine responder children demonstrate distinct mononuclear cell DNA methylation pathways. *Clin Epigenetics.* 2024;16:85.
 42. Bakulski KM, Feinberg JI, Andrews SV, Yang J, Brown S, S LM, Witter F, Walston J, Feinberg AP, Fallin MD. DNA methylation of cord blood cell types: applications for mixed cell birth studies. *Epigenetics.* 2016;11:354–362.
 43. de Goede OM, Razzaghi HR, Price EM, Jones MJ, Kobor MS, Robinson WP, Lavoie PM. Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. *Clin Epigenetics.* 2015;7:95.
 44. De Goede OM, Razzaghi HR, Price EM, Jones MJ, Kobor MS, Robinson WP, Lavoie PM. Nucleated red blood cells impact DNA methylation and expression analyses of cord blood hematopoietic cells. 2015. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68456>.
 45. Lin X, Tan JYL, Teh AL, Lim IY, Liew SJ, MacIsaac JL, Chong YS, Gluckman PD, Kobor MS, Cheong CY, Karnani N. Cell type-specific DNA methylation in neonatal cord tissue and cord blood: a 850K-reference panel and comparison of cell types. *Epigenetics.* 2018;13:941–58.
 46. De Goede OM, Lavoie PM, Robinson WP. Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. *Clinical Epigenetics.* 2017;9:1.

47. De Goede OM, Lavoie PM, Robinson WP: Cord blood hematopoietic cells from preterm infants display altered DNA methylation patterns. 2017. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE82084>.
48. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC: Enhanced cell deconvolution of peripheral blood using DNA methylation for high-resolution immune profiling. 2022. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE180970>.
49. Gervin K, Salas LA, Bakulski KM, van Zelm MC, Koestler DC, Wiencke JK, Duijts L, Moll HA, Kelsey KT, Kobor MS, et al.: Systematic evaluation and validation of reference and library selection methods for deconvolution of cord blood DNA methylation data. 2019. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE127824>.
50. Monteagudo-Sánchez A, Hernandez Mora JR, Simon C, Burton A, Tenorio J, Lapunzina P, Clark S, Esteller M, Kelsey G, López-Siguero JP, et al.: The role of ZFP57 and additional KRAB-zinc finger proteins in the maintenance of human imprinted methylation and multi-locus imprinting disturbances. *Nucleic Acids Res.* 2020;48:11394–407.
51. Monteagudo-Sánchez A, Hernandez Mora JR, Simon C, Burton A, Tenorio J, Lapunzina P, Clark S, Esteller M, Kelsey G, López-Siguero JP, et al.: The role of ZFP57 and additional KRAB-zinc finger proteins in the maintenance of human imprinted methylation and multi-locus imprinting disturbances. 2020. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149572>.
52. Puvvula J, Braun JM, DeFranco EA, Ho S-M, Leung Y-K, Huang S, Zhang X, Vuong AM, Kim SS, Percy Z, et al.: Gestational exposure to environmental chemicals and epigenetic alterations in the placenta and cord blood mononuclear cells. *Epigenetics Commun.* 2024;4:4.
53. Puvvula J, Braun JM, DeFranco EA, Ho SM, Leung YK, Huang S, Zhang X, Vuong AM, Kim SS, Percy Z, et al.: Gestational exposure to environmental chemicals and epigenetic alterations in the placenta and cord blood mononuclear cells. 2024. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE269983>.
54. Solomon O, Yousefi P, Huen K, Gunier RB, Escudero-Fung M, Barcellos LF, Eskenazi B, Holland N: Prenatal phthalate exposure and altered patterns of DNA methylation in cord blood. *Environ Mol Mutagen.* 2017;58:398–410.
55. Solomon O, Yousefi P, Huen K, Gunier RB, Escudero-Fung M, Barcellos LF, Eskenazi B, Holland N: Prenatal phthalate exposure and altered patterns of DNA methylation in cord blood. 2017. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE97628>.
56. Wang X, Cho H-Y, Campbell MR, Panduri V, Coviello S, Caballero MT, Sambandan D, Kleeberger SR, Polack FP, Ofman G, Bell DA: Epigenome-wide association study of bronchopulmonary dysplasia in preterm infants: results from the discovery-BPD program. *Clin Epigenetics.* 2022;14:57.
57. Wang X, Cho HY, Campbell MR, Panduri V, Coviello S, Caballero MT, Sambandan D, Kleeberger SR, Polack FP, Ofman G, Bell DA: Epigenome-wide association study of bronchopulmonary dysplasia in preterm infants: results from the discovery-BPD program. 2022. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE188949>.
58. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST: Age-associated DNA methylation in pediatric populations. *Genome Res.* 2012;22:623–32.
59. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST: Age-associated DNA methylation in pediatric populations. 2012. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE36064>.
60. Quinn EB, Hsiao CJ, Maisha FM, Mulligan CJ: Low birthweight is associated with epigenetic age acceleration in the first 3 years of life. *Evol Med Public Health.* 2023;11:251–61.
61. Quinn EB, Hsiao CJ, Maisha FM, Mulligan CJ: Low birthweight is associated with epigenetic age acceleration in the first 3 years of life. 2023. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE224573>.
62. Shi L, Jiang F, Ouyang F, Zhang J, Wang Z, Shen X: DNA methylation markers in combination with skeletal and dental ages to improve age estimation in children. *Forensic Sci Int Genet.* 2018;33:1–9.
63. Shi L, Jiang F, Ouyang F, Zhang J, Wang Z, Shen X: DNA methylation markers in combination with skeletal and dental ages to improve age estimation in children. 2018. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE104812>.
64. Wang X, Campbell MR, Cho H-Y, Pittman GS, Martos SN, Bell DA: Epigenomic profiling of isolated blood cell types reveals highly specific B cell smoking signatures and links to disease risk. *Clin Epigenetics.* 2023;15:90.
65. Wang X, Campbell MR, Cho HY, Pittman GS, Martos SN, Bell DA: Epigenomic profiling of isolated blood cell types reveals highly specific B cell smoking signatures and links to disease risk. 2023. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE224807>.
66. Salas LA, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC: An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray. 2018. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112618>.
67. Salas LA, Zhang Z, Koestler DC, Butler RA, Hansen HM, Molinaro AM, Wiencke JK, Kelsey KT, Christensen BC: FlowSorted.BloodExtended.EPIC: an enhanced DNA methylation library for deconvoluting peripheral blood. 2022. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE167998>.
68. Horvath S: DNA methylation age of human tissues and cell types. 2013. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE56581>.
69. Stunnenberg HG, International Human Epigenome C, Hirst M: The international human epigenome consortium: a blueprint for scientific collaboration and discovery (BLUEPRINT data). 2016. <https://ega-archive.org/datasets/EGAD00010000850>.
70. Zillbauer MR, TF Clark, C Coffey, AJ Joyce, CJ Palta, P Palotie, A Lyons, PA Smith, KG: Genome-wide methylation analyses of primary human leukocyte subsets identifies functionally important cell-type-specific hypomethylated regions. 2012. <https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-2145>.
71. Coit P, Yalavarthi S, Ognenovski M, Zhao W: Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. 2016. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE65097>.
72. Massanella M, Tusell CM, Pujantell M, Cap P, Rallón N, García-García M, Carrasco A, Restrepo-Montoya D, Company C, Coll J, et al.: The T-cell DNA methylome captures part of the host-pathogen interplay in HIV-1 infection. 2015. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50222>.
73. Shade DP, HJ Hausman, DB Hohos, N Meagher, RB Kauwell, GP Lewis, RD Smith, AK Bailey, LB: DNA methylation changes in whole blood and CD16+ neutrophils in response to chronic folic acid supplementation in women of childbearing age. 2016. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63499>.
74. Limbach M, Saare M, Tserel L, Kisand K: Epigenetic profiling in CD4+ and CD8+ T cells from Graves' disease patients reveals changes in genes associated with T cell receptor signaling. 2016. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71955>.
75. Mamrut SA, N Staun-Ram E, Ginzburg E: Integrative analysis of methylome and transcriptome in human blood identifies extensive sex- and immune cell-specific differentially methylated regions. 2012. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE71244>.
76. Absher DL, X Waite, LL Gibson, A: Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. 2016. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59250>.
77. Tserel LK, R Limbach, M Tretyakov K: Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. 2016. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59065>.
78. Paul DS, Teschendorff AE, Dang MA, Lowe R, Hawa MI, Ecker S, Beyan H, Cunningham S, Fouts AR, Ramelius A, et al.: Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. 2016. <https://www.ebi.ac.uk/ega/studies/EGAS00001001598>.
79. Reinis LE, Acevedo N, Joerink M, Pershagen G, Dahlén SE, Greco D, Söderhäll C, Scheynius A, Kere J: Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. 2012. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE35069>.

80. Shang LZ, Wang W, Li Z: meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. 2023. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210254>.
81. Zannas AS, Jia M, Hafner K, Baumert J, Wiechmann T, Pape JC, Arloth J, Kodel M, Martinelli S, Roitman M, et al.: Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-kappaB-driven inflammation and cardiovascular risk. 2019. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72680>.
82. Johansson A, Enroth S, Gyllenstein U: Continuous aging of the human DNA methylome throughout the human lifespan. 2013. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87571>.
83. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, Klotzle B, Bibikova M, Fan JB, Gao Y, et al.: Genome-wide methylation profiles reveal quantitative views of human aging rates. 2013. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279>.
84. Bujold D, Morais DAdL, Gauthier C, Côté C, Caron M, Kwan T, Chen KC, Laperle J, Markovits AN, Pastinen T, et al.: The international human epigenome consortium data portal. 2016. <https://epigenomesportal.ca/ihec/>.
85. Lorente-Pozo S, Navarrete P, Garzon MJ, Lara-Canton I, Beltran-Garcia J, Osca-Verdegai R, Mena-Molla S, Garcia-Lopez E, Vento M, Pallardo FV, Garcia-Gimenez JL: DNA methylation analysis to unravel altered genetic pathways underlying early onset and late onset neonatal sepsis. A pilot study. *Front Immunol.* 2021;12:622599.
86. Lorente-Pozo S, Navarrete P, Garzon MJ, Lara-Canton I, Beltran-Garcia J, Osca-Verdegai R, Mena-Molla S, Garcia-Lopez E, Vento M, Pallardo FV, Garcia-Gimenez JL: DNA methylation analysis to unravel altered genetic pathways underlying early onset and late onset neonatal sepsis. A pilot study. 2021. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE155952>.
87. Henneman P, Bouman A, Mul A, Knecht L, van der Kevie-Kersemakers AM, Zwaveling-Soonawala N, Meijers-Heijboer HEJ, van Trotsenburg ASP, Mannens MM: Widespread domain-like perturbations of DNA methylation in whole blood of Down syndrome neonates. *PLoS ONE.* 2018;13: e0194938.
88. Henneman P, Bouman A, Mul A, Knecht L, van der Kevie-Kersemakers AM, Zwaveling-Soonawala N, Meijers-Heijboer HEJ, van Trotsenburg ASP, Mannens MM: Widespread domain-like perturbations of DNA methylation in whole blood of Down syndrome neonates. 2018. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107211>.
89. Naumova OY, Lipschutz R, Rychkov SY, Zhukova OV, Grigorenko EL: DNA methylation alterations in blood cells of toddlers with down syndrome. *Genes (Basel).* 2021;12:12.
90. Naumova OY, Lipschutz R, Rychkov SY, Zhukova OV, Grigorenko EL: DNA methylation alterations in blood cells of toddlers with down syndrome. 2021. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE174555>.
91. Bacalini MG, Gentilini D, Boattini A, Giampieri E, Pirazzini C, Giuliani C, Fontanesi E, Scurti M, Remondini D, Capri M, et al.: Identification of a DNA methylation signature in blood cells from persons with down syndrome. *Aging (Albany NY).* 2015;7:82–96.
92. Bacalini MG, Gentilini D, Boattini A, Giampieri E, Pirazzini C, Giuliani C, Fontanesi E, Scurti M, Remondini D, Capri M, et al.: Identification of a DNA methylation signature in blood cells from persons with down syndrome. 2015. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE52588>.
93. Chen KD, Huang YH, Guo MMH, Chang LS, Chu CH, Bu LF, Chu CL, Lee CH, Liu SF, Kuo HC: DNA methylation array identifies golli-MBP as a biomarker for disease severity in childhood atopic dermatitis. *J Invest Dermatol.* 2022;142:104–13.
94. Chen KD, Huang YH, Guo MH, Chang LS, Chu CH, Bu LF, Chu CL, Lee CH, Liu SF, Kuo HC: DNA methylation array identifies Golli-MBP as a biomarker for disease severity in childhood atopic dermatitis. 2022. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152084>.
95. Yang IV, Pedersen BS, Liu A, O'Connor GT, Teach SJ, Kattan M, Misiak RT, Gruchalla R, Steinbach SF, Szefer SJ, et al.: DNA methylation and childhood asthma in the inner city. *J Allergy Clin Immunol.* 2015;136:69–80.
96. Yang IV, Pedersen BS, Liu A, O'Connor GT, Teach SJ, Kattan M, Misiak RT, Gruchalla R, Steinbach SF, Szefer SJ, et al.: DNA methylation and childhood asthma in the inner city. 2015. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40576>.
97. Langie SAS, Moisse M, Szarc Vel Szic K, Van Der Plas E, Koppen G, De Prins S, Louwies T, Nelen V, Van Camp G, Lambrechts D, et al.: GLI2 promoter hypermethylation in saliva of children with a respiratory allergy. *Clin Epigenetics.* 2018;10:50.
98. Langie SAS, Moisse M, Szarc Vel Szic K, Van Der Plas E, Koppen G, De Prins S, Louwies T, Nelen V, Van Camp G, Lambrechts D, et al.: GLI2 promoter hypermethylation in saliva of children with a respiratory allergy. 2018. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE110128>.
99. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, Shorosh H, Fingerlin T, Frohnert BI, Yang IV, et al.: Longitudinal DNA methylation differences precede type 1 diabetes. *Sci Rep.* 2020;10:3721.
100. Johnson RK, Vanderlinden LA, Dong F, Carry PM, Seifert J, Waugh K, Shorosh H, Fingerlin T, Frohnert BI, Yang IV, et al.: Longitudinal DNA methylation differences precede type 1 diabetes. 2020. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE142512>.
101. Paul DS, Teschendorff AE, Dang MA, Lowe R, Hawa MI, Ecker S, Beyan H, Cunningham S, Fouts AR, Ramelius A, et al.: Increased DNA methylation variability in type 1 diabetes across three immune effector cell types. *Nat Commun.* 2016;7:13555.
102. Smyth GK: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol.* 2004;3:Article3.
103. Zhuang J, Widschwendter M, Teschendorff AE: A comparison of feature selection and classification methods in DNA methylation studies using the Illumina Infinium platform. *BMC Bioinformatics.* 2012;13: 59.
104. Zannas AS, Jia M, Hafner K, Baumert J, Wiechmann T, Pape JC, Arloth J, Kodel M, Martinelli S, Roitman M, et al.: Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-kappaB-driven inflammation and cardiovascular risk. *Proc Natl Acad Sci U S A.* 2019;116:11370–9.
105. Johansson A, Enroth S, Gyllenstein U: Continuous aging of the human DNA methylome throughout the human lifespan. *PLoS ONE.* 2013;8: e67378.
106. You C, Wu S, Zheng SC, Zhu T, Jing H, Flagg K, Wang G, Jin L, Wang S, Teschendorff AE: A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. *Nat Commun.* 2020;11:4779.
107. Peng Q, Liu X, Li W, Jing H, Li J, Gao X, Luo Q, Breeze CE, Pan S, Zheng Q, et al.: Analysis of blood methylation quantitative trait loci in East Asians reveals ancestry-specific impacts on complex traits. *Nat Genet.* 2024;56:846–60.
108. Zheng SC, Breeze CE, Beck S, Teschendorff AE: Identification of differentially methylated cell types in epigenome-wide association studies. *Nat Methods.* 2018;15:1059–66.
109. Goronzy JJ, Fang F, Cavanagh MM, Qi Q, Weyand CM: Naive T cell maintenance and function in human aging. *J Immunol.* 2015;194:4073–80.
110. Moskowitz DM, Zhang DW, Hu B, Le Saux S, Yanes RE, Ye Z, Buenrostro JD, Weyand CM, Greenleaf WJ, Goronzy JJ: Epigenomics of human CD8 T cell differentiation and aging. *Sci Immunol.* 2017;2:eaag0192.
111. Jonkman TH, Dekkers KF, Slieker RC, Grant CD, Ikram MA, van Greevenbroek MMJ, Franke L, Veldink JH, Boomsma DI, Slagboom PE, et al.: Functional genomics analysis identifies T and NK cell activation as a driver of epigenetic clock progression. *Genome Biol.* 2022;23:24.
112. Malle L, Patel RS, Martin-Fernandez M, Stewart OJ, Philippot Q, Buta S, Richardson A, Barcessat V, Taft J, Bastard P, et al.: Autoimmunity in Down's syndrome via cytokines, CD4 T cells and CD11c(+) B cells. *Nature.* 2023;615:305–14.
113. Henry E, Walker D, Wiedmeier SE, Christensen RD: Hematological abnormalities during the first week of life among neonates with down syndrome: data from a multihospital healthcare system. *Am J Med Genet A.* 2007;143A:42–50.
114. Szczawinska-Poplonyk A, Poplonyk N, Awdi K: Down syndrome in children: a primary immunodeficiency with immune dysregulation. *Children (Basel).* 2024;11:11.
115. Versteegen RH, Kusters MA, Gemen EF, de Vries E: Down syndrome B-lymphocyte subpopulations, intrinsic defect or decreased T-lymphocyte help. *Pediatr Res.* 2010;67(5):563–9.
116. Huggard D, Doherty DG, Molloy EJ: Immune Dysregulation in Children With Down Syndrome. *Front Pediatr.* 2020;8: 73.

117. Dulay AT, Buhimschi IA, Zhao G, Luo G, Abdel-Razek S, Cackovic M, Rosenberg VA, Pettker CM, Thung SF, Bahtiyar MO, et al. Nucleated red blood cells are a direct response to mediators of inflammation in newborns with early-onset neonatal sepsis. *Am J Obstet Gynecol*. 2008;198(426):e421–429.
118. Sumitro KR, Utomo MT, Widodo ADW. Neutrophil-to-lymphocyte ratio as an alternative marker of neonatal sepsis in developing countries. *Oman Med J*. 2021;36: e214.
119. Gerday S, Schleich F, Henket M, Guissard F, Paulus V, Louis R. Revisiting differences between atopic and non-atopic asthmatics: when age is shaping airway inflammatory profile. *World Allergy Organ J*. 2022;15: 100655.
120. Long H, Zhang G, Wang L, Lu Q. Eosinophilic skin diseases: a comprehensive review. *Clin Rev Allergy Immunol*. 2016;50:189–213.
121. Lambrecht BN, Hammad H. The immunology of asthma. *Nat Immunol*. 2015;16:45–56.
122. Hu Y, Liu S, Liu P, Mu Z, Zhang J. Clinical relevance of eosinophils, basophils, serum total IgE level, allergen-specific IgE, and clinical features in atopic dermatitis. *J Clin Lab Anal*. 2020;34: e23214.
123. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. *Bioinformatics*. 2016;32:286–8.
124. Herold KC, Delong T, Perdigoto AL, Biru N, Brusko TM, Walker LSK. The immunology of type 1 diabetes. *Nat Rev Immunol*. 2024;24:435–51.
125. Noguera-Castells A, Garcia-Prieto CA, Alvarez-Errico D, Esteller M. Validation of the new EPIC DNA methylation microarray (900K EPIC v2) for high-throughput profiling of the human DNA methylome. *Epigenetics*. 2023;18: 2185742.
126. Goldberg DC, Cloud C, Lee-Moe S, Barnes B, Gruber S, Kim E, Pottekat A, Westphal MS, McAuliffe L, Majounie E, et al. Scalable screening of ternary-code DNA methylation dynamics associated with human traits. *bioRxiv*. 2024. <https://doi.org/10.1101/2024.05.17.594606>.
127. Breeze CE, Beck S, Berndt SI, Franceschini N. The missing diversity in human epigenomic studies. *Nat Genet*. 2022;54:737–9.
128. Breeze CE, Wong JYY, Beck S, Berndt SI, Franceschini N. Diversity in EWAS: current state, challenges, and solutions. *Genome Med*. 2022;14:71.
129. Min JL, Hemani G, Hannon E, Dekkers KF, Castillo-Fernandez J, Luijk R, Carnero-Montoro E, Lawson DJ, Burrows K, Suderman M, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nat Genet*. 2021;53:1311–21.
130. Chen Y, Wu Z, Sutlive J, Wu K, Mao L, Nie J, Zhao XZ, Guo F, Chen Z, Huang Q. Noninvasive prenatal diagnosis targeting fetal nucleated red blood cells. *J Nanobiotechnology*. 2022;20:546.
131. Zheng SC, Webster AP, Dong D, Feber A, Graham DG, Sullivan R, Jevons S, Lovat LB, Beck S, Widschwendter M, Teschendorff AE. A novel cell-type deconvolution algorithm reveals substantial contamination by immune cells in saliva, buccal and cervix. *Epigenomics*. 2018;10:925–40.
132. Osumi-Sutherland D, Xu C, Keays M, Levine AP, Kharchenko PV, Regev A, Lein E, Teichmann SA. Cell type ontologies of the human cell atlas. *Nat Cell Biol*. 2021;23:1129–35.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.