A meta-analysis of immune-cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes

Qi Luo[1†], Varun B. Dwaraka[2†], Qingwen Chen[3†], Huige Tong[1], Tianyu Zhu[1], Kirsten Seale[4], Joseph M. Raffaele[5], Shijie C. Zheng[6], Tavis L. Mendez[2], Yulu Chen[3], Natalia Carreras[2], Sofina Begum[3], Kevin Mendez[3], Sarah Voisin[4], Nir Eynon[7], Jessica A. Lasky-Su[3*], Ryan Smith[2*] and Andrew E. Teschendorff[1*]

Abstract

Background Changes in cell-type composition of tissues are associated with a wide range of diseases and environmental risk factors and may be causally implicated in disease development and progression. However, these shifts in cell-type fractions are often of a low magnitude, or involve similar cell subtypes, making their reliable identification challenging. DNA methylation profiling in a tissue like blood is a promising approach to discover shifts in cell-type abundance, yet studies have only been performed at a relatively low cellular resolution and in isolation, limiting their power to detect shifts in tissue composition.

Methods Here we derive a DNA methylation reference matrix for 12 immune-cell types in human blood and extensively validate it with flow-cytometric count data and in whole-genome bisulfite sequencing data of sorted cells. Using this reference matrix, we perform a directional Stouffer and fixed effects meta-analysis comprising 23,053 blood samples from 22 different cohorts, to comprehensively map associations between the 12 immune-cell fractions and common phenotypes. In a separate cohort of 4386 blood samples, we assess associations between immune-cell fractions and health outcomes.

Results Our meta-analysis reveals many associations of cell-type fractions with age, sex, smoking and obesity, many of which we validate with single-cell RNA sequencing. We discover that naïve and regulatory T-cell subsets are higher in women compared to men, while the reverse is true for monocyte, natural killer, basophil, and eosinophil fractions. Decreased natural killer counts associated with smoking, obesity, and stress levels, while an increased count correlates with exercise and sleep. Analysis of health outcomes revealed that increased naïve CD4 + T-cell and N-cell fractions associated with a reduced risk of all-cause mortality independently of all major epidemiological risk factors and baseline co-morbidity. A machine learning predictor built only with immune-cell fractions achieved a C-index value

---

†Qi Luo, Varun B. Dwaraka and Qingwen Chen contributed equally to this work.

*Correspondence:
Jessica A. Lasky-Su
rejas@channing.harvard.edu
Ryan Smith
ryan@trudiagnostic.com
Andrew E. Teschendorff
andrew@sinh.ac.cn
Full list of author information is available at the end of the article

for all-cause mortality of 0.69 (95%CI 0.67–0.72), which increased to 0.83 (0.80–0.86) upon inclusion of epidemiological risk factors and baseline co-morbidity.

Conclusions This work contributes an extensively validated high-resolution DNAm reference matrix for blood, which is made freely available, and uses it to generate a comprehensive map of associations between immune-cell fractions and common phenotypes, including health outcomes.

Keywords Immune system, Disease risk factors, Aging, Sex, Obesity, Epigenetic clocks, Mortality, Covid-19, Cancer

Background

Human tissues contain many different cell types in proportions that vary between healthy individuals as well as in association with disease and exposure to disease risk factors [1, 2]. These shifts in cell-type proportions may not only constitute important biomarkers of environmental exposures, disease risk, or early diagnosis, but may be causally implicated, as exemplified by immune-cell variations that impact cancer progression [3] and immunosenescence [4]. Although detecting shifts in cell-type composition in easily accessible tissues like blood has been possible with moderately sized studies in the context of autoimmune diseases, cancer, or aging [5–8], detecting more subtle changes in cell-type proportions that may arise in relation to disease risk factors like sex, obesity, or smoking, has been more challenging. This is not only because underlying shifts in cell-type proportions may be of low magnitude, typically involving only a few percentage points, but also because there is already substantial variation in these proportions between healthy or unexposed individuals. Thus, measuring cell counts in large cohorts of samples is necessary in order to confidently identify disease-or-exposure-associated shifts in cell-type composition. However, experimental cell-counting methods are cumbersome and not easily scalable to thousands of samples.

DNA methylation (DNAm) has been abundantly profiled in easily accessible and heterogeneous tissues like whole blood [9–15], saliva [16, 17], and buccal swabs [18]. The underlying cell-type heterogeneity (CTH) of these tissues thus offers the opportunity to detect phenotype-associated shifts in cell-type composition. Indeed, because DNAm is highly cell-type specific and can be measured with high accuracy [19], application of cell-type deconvolution algorithms [2, 20] to average DNAm profiles generated by epigenome-wide association studies (EWAS) has proved to be an excellent means to accurately quantify the underlying cell-type fractions in a wide range of complex tissues [20–23]. Not until recently, the main limitation has been the availability of a high-resolution tissue-specific DNAm reference matrix, containing representative DNAm profiles for all cell types in the tissue of interest and which is required by

reference-based cell-type deconvolution methods to infer the underlying cell-type fractions [2, 22–27].

Here we use the Illumina 850k DNAm profiles of cell-sorted samples from Salas et al. [28] to build a novel 12 immune-cell-type DNAm reference matrix for blood tissue, using an improved procedure that exclusively uses cell-type-specific unmethylated markers [29]. We validate this 12 immune-cell-type DNAm reference matrix on DNAm data with matched flow-cytometric counts, as well as in a large collection of immune-cell-sorted samples, including whole-genome-bisulfite sequencing (WGBS) samples from the International Human Epigenome Consortium (IHEC) [30]. Importantly, we collate genome-wide DNAm data for a total of 22 independent cohorts, encompassing over 23,000 blood samples, and use our DNAm reference matrix to perform a meta-analysis of immune-cell-type fraction associations with common phenotypes, including age, sex, smoking, and obesity, validating or strengthening previous findings, while also revealing novel associations. In a large cohort with extensive epidemiological and health outcome annotation for approximately 4000 to 6000 whole blood samples, we identify additional associations of immune-cell fractions with exercise, alcohol consumption, stress, and health outcomes, including all-cause mortality. In summary, we use a high-quality high-resolution DNAm reference matrix to comprehensively map associations of 12 immune-cell-type fractions with common phenotypes and health outcomes.

Methods

Construction of the 12 immune-cell-type DNAm reference matrices

We obtained the EPIC DNAm dataset of 12 sorted immune-cell subsets from GEO under accession number GSE167988. The idat files were downloaded and processed using *minfi* R-package [31]. We retained 756,625 probes with significantly detected values across all 68 samples. The resulting beta-valued data matrix was then adjusted for type-2 probe bias using BMIQ [32]. We next removed 12 artificial mixture samples, leaving a total of 56 sorted samples: 6 basophils (Baso), 6 memory B-cells (Bmem), 4 naïve B-cells (Bnv), 4 memory CD4 + T-cells (CD4Tmem), 5 naïve CD4 + T-cells

Luo *et al. Genome Medicine*   (2023) 15:59

Page 3 of 24

(CD4Tnv), 4 memory CD8 + T-cells (CD8Tmem), 5 naïve CD8 + T-cells (CD8Tnv), 4 eosinophils (Eos), 5 monocytes (Mono), 6 neutrophils (Neu), 4 natural killer (NK), and 3 regulatory T-cells (Treg). We then performed SVD, estimating the number of significant components using RMT [33], followed by hierarchical clustering on the sample projection matrix, to check that sorted samples clustered by cell type. This revealed that one CD4Tmem and one CD8Tnv sample did not cluster correctly. Hence, these two samples were removed leaving a normalized beta-valued data matrix over 756,625 probes and 54 sorted samples. We then used *limma* [34, 35] to perform differential DNAm analysis for each of the 12 cell types in turn, comparing that cell type to the other eleven. Next, for each cell type, we selected those probes passing a Benjamini–Hochberg adjusted FDR < 0.001 and displaying hypomethylation in the given cell-type compared to the rest. We only consider hypomethylated probes because these are overwhelmingly more likely to be truly cell-type-specific markers [24, 29]. The above procedure can still result in probes attaining smaller DNAm values in another cell type because the limma analysis compares an average of the given cell type to the average over 11 other cell types. Hence, to ensure that our selected probes for a given cell type attain the smallest DNAm values in that cell type, we also recorded for each cell type t and probe p the minimum DNAm difference value (called Δ_{pt}) between the cell type of interest t and the other 11 cell types. We note that for the probes of interest, these values will be negative because the maximum value across the samples of the given cell type of interest should be lower than the minimum value across all other cell types. Then for all hypomethylated probes at FDR < 0.001 for a given cell type t, we ranked these in increasing order of the Δ_{pt} values, to ensure that the top-ranked probes display the largest negative Δ_{pt} values. This means that for these probes the maximum beta DNAm value across the samples of the given cell type of interest is much smaller than the minimum value across the samples from all other cell types, which ensures that we are selecting probes with the largest effect sizes. For each cell type, we then selected the top-ranked 50 probes as cell-type-specific markers, resulting in a total of 600 (50 times 12) unique marker probes. We note that the number of hypomethylated probes at FDR < 0.001 per cell type was in general quite large: mean over the 12 cell types was 24,954, range was 1111 (CD4Tmem) to 88,992 (Bmem). Hence, selecting the top-ranked 50 according to Δ_{pt} values ensures that we are selecting not only highly significant hypomethylated probes but also those with the largest possible effect sizes. The final DNAm reference matrix over the 600 marker probes was then built by taking the median DNAm value over the samples of

a given cell type. Note that we take the median, because this is a more robust estimator and because later we estimate cell-type fractions using a robust partial correlation framework which does not require the assumption that the reference value should be an average (in contrast to constrained projection which does). In order not to bias performance in Illumina 450k datasets, we also generated a separate 12 cell-type DNAm reference using only 450k probes, using the exact same procedure as described above. Of note this 450k DNAm reference matrix is also defined for 600 unique marker probes.

Validation DNA methylation datasets of sorted samples
We obtained independent immune-cell-sorted samples from the following sources, encompassing both Illumina 450k and WGBS technologies: From Reynolds et al. [36], we obtained 1202 monocyte and 214 CD4 + T-cell 450k samples (GEO: GSE56581). From BLUEPRINT [37], we obtained 139 monocyte, 139 naïve CD4 + T-cell, and 139 neutrophil 450k samples from the same 139 individuals. From Zilbauer et al., we obtained 6 CD4 + T-cell, 6 CD8 + T-cell, 6 B-cell, 6 neutrophil, and 6 monocyte 450k samples (ArrayExpress: E-MTAB-2145). From Coit et al. [38], we obtained 15 neutrophil 450k samples (GEO: GSE65097). From Nestor et al. [39], we obtained 8 CD4 + T-cell 450k samples (GEO: GSE50222). From Shade et al. [40], we obtained 12 neutrophil 450k samples (GEO: GSE63499). From Limbach et al. [41], we obtained 31 CD4 + T-cell and 31 CD8 + T-cell 450k samples (GEO: GSE71955). From Mamrut et al. [42], we obtained 6 CD4 + T-cell, 5 CD8 + T-cell, 4 B-cell, and 5 monocyte 450k samples (GEO: GSE71244). From Absher et al. [43], we obtained 71 CD4 + T-cell, 56 B-cell, and 28 monocyte 450 k samples (GEO: GSE59250). From Tserel et al. [44], we obtained 99 CD4 + T-cell and 100 CD8 + T-cell 450k samples (GEO: GSE59065). From Paul et al. [45], we obtained 49 CD4 + T-cell, 50 B-cell, and 52 monocyte 450k samples (EGA: EGAS00001001598). From Reinius et al. [46], we obtained 6 CD4 + T-cell, 6 CD8 + T-cell, 6 B-cell, 6 neutrophil, 6 monocyte, and 6 eosinophil 450k samples. From the IHEC data portal (https://epigenomes portal.ca/ihec/), we obtained WGBS hg38 immune-cell-sorted samples from build version 2020–10. The downloaded files were in bigwig format. For each sample, there are 2 bigwig files, one for read coverage information and the other for beta value information. We first used bigWigToWig shell script provided by UCSC genome browser to convert them into wig files. Then for each sample, we combined the read coverage and beta value information into one file and stored them as an.rda file for further processing. For each WGBS sample, we found the CpGs present in the 850k DNAm beadarray. We dropped one sample (ERS568736) due to ultra-low coverage. For

Luo *et al. Genome Medicine*   (2023) 15:59

Page 4 of 24

the rest of samples, the minimum number of 850k probes covered by 20 reads or more was 336,812, the maximum was 834,149, with a mean of 689,451. In total, for this work, we used 4 memory CD4 + T-cell, 2 T-regulatory, 4 naïve CD8 + T-cell, 2 memory CD8 + T-cell, 7 naïve B-cell, 5 memory B-cell, 12 neutrophil, 22 monocyte, 2 eosinophil, and 4 natural killer cell WGBS samples, for validating our 850 DNAm reference matrix.

In silico mixture validation analysis with WGBS cell-sorted samples

The overall coverage of common 850k probes across all WGBS samples was only 132,713, containing only 72 probes from our 600 CpG 850k DNAm reference matrix. Hence, for the in silico mixture analysis, we did not impose any threshold on read coverage, which resulted in 487,795 probes, including 304 probes from our 850k DNAm reference matrix. This is sensible because reference-based cell-type deconvolution can tolerate even up to 30% errors in the DNAm reference matrix [47]. Hence, in silico mixtures were generated from the 4 memory CD4 + T-cell, 2 T-regulatory, 4 naïve CD8 + T-cell, 2 memory CD8 + T-cell, 7 naïve B-cell, 5 memory B-cell, 12 neutrophil, 22 monocyte, 2 eosinophil, and 4 natural killer cell WGBS samples, defined over the 304 probes. We generated 1000 in silico mixtures, randomly selecting one sample from each immune-cell type and using random weights drawn from a uniform distribution to generate the linear combination. Since there are a total of 4*2*4*2*7*5* 12*22*2*4 = 4,730,880 potential combinations, generating 1000 in silico mixtures is sensible as the number is large enough to reliably assess performance, while also reducing statistical dependency of the combinations as much as possible. Performance was assessed using Pearson R-values and RMSE.

Illumina DNA methylation datasets used in the meta-analysis

Below we provide details of the data source and processing of each dataset used in our meta-analyses. Each dataset profiled whole or peripheral blood samples with Illumina DNAm beadarrays (EPIC or 450k). Further details are available in Additional File 1: table S3. In all cases where idat files were available, we processed the data with a uniform procedure that used *minfi* for processing with Illumina normalization method [31], followed by BMIQ normalization to correct for type-2 probe bias [32]. This strategy to normalize the DNAm data with *minfi* followed by BMIQ normalization has been shown to work reasonably well [48–51].

LiuMS

The 450k dataset from Kular et al. [52] was obtained from the NCBI GEO website under the accession number GSE106648 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106648). We downloaded the series matrix file which contained the data processed with detection *P*-values. We only retained probes with no missing values across all samples. This data was subsequently normalized with BMIQ [32], resulting in a normalized data matrix for 483,567 probes and 279 peripheral blood samples (140 multiple sclerosis patients + 139 controls).

Song

The EPIC dataset from Song et al. [53] profiled DNAm in blood from childhood cancer survivors and was obtained from the NCBI GEO website under the accession number GSE169156 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169156). The file "GSE169156_RAW.tar" which contains the IDAT files was downloaded and processed with *minfi* package [31]. Probes with *P*-values < 0.05 across all samples were retained. The filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 823,395 probes and 2052 samples.

HPT-EPIC & HPT-450k

These datasets derived DNAm profiles from the peripheral blood of African-Americans as part of The Genetic Epidemiology Network of Arteriopathy (GENOA) study [54]. Data was obtained from the NCBI GEO websites under the accession numbers GSE210255 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210255, Infinium HumanMethylationEPIC BeadChip) and GSE210254 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210254, Infinium HumanMethylation450k BeadChip). The files "GSE210255_RAW.tar" and "GSE210254_RAW.tar" containing the IDAT files were downloaded and processed with minfi package. Probes with *P*-values < 0.05 across all samples were retained. The filtered data was subsequently normalized with BMIQ, resulting in normalized data matrices containing 826,512 probes and 1394 samples (EPIC set), and 476,722 probes and 418 samples (450k set), respectively.

Barturen

The EPIC dataset from Barturen et al. [55] profiled DNAm in blood from Covid-19 patients with three different levels of disease severity. Data was obtained from the NCBI GEO website under accession number GSE179325 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179325). The file "GSE179325_RAW.tar" containing IDAT files was downloaded and processed with minfi package. Only probes with *P*-values < 0.05

Luo *et al. Genome Medicine*   (2023) 15:59

Page 5 of 24

across all samples were retained. The filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 845,921 probes and 574 samples.

Airwave

The EPIC dataset from the Airwave study [56] was obtained from the NCBI GEO website under accession number GSE147740 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147740). The file "GSE147740_RAW.tar" containing IDAT files was downloaded and processed with minfi. Only probes with P-values < 0.05 across all samples were retained. Filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 840,034 probes and 1129 samples.

VACS

The 450k dataset from Zhang X et al. [57] was obtained from the NCBI GEO website under accession number GSE117860 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117860). The file "GSE117860_MethylatedSignal.txt.gz" containing the unmethylated signals, methylated signals, and detection P-values was downloaded. The beta values were obtained using the Illumina definition: methylated signal / (methylated signal + unmethylated signal + 100). Only probes with P-values < 0.05 across all samples were retained. The filtered beta value matrix was subsequently normalized with BMIQ, resulting in a normalized data matrix 396,327 probes across 529 samples.

Ventham

The 450k dataset from Ventham et al. [58] was obtained from NCBI GEO website under accession number GSE87648 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87648). The file "GSE87648_RAW.tar" containing the IDAT files was downloaded and processed with minfi package. Two samples in which the proportion of probes with P-values < 0.05 is lower than 0.99 were excluded, and probes with P-values < 0.05 across all remaining samples were kept. Filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 470,807 probes and 382 samples.

Hannon-1 and 2

The 450k datasets from Hannon et al. [59, 60] were obtained from NCBI GEO websites under accession numbers GSE80417 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80417) and GSE84727 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84727), which represent the phase-1 and phase-2 of their study. For the phase-1

dataset, the file "GSE80417_rawBetas.csv.gz" containing the beta values for filtered probes was downloaded, and the beta values were then normalized with BMIQ algorithm. For the phase-2 dataset, the file "GSE84727_rawBetas.csv.gz" and "GSE84727_detectionP.csv.gz" were downloaded. Only probes with P-values < 0.05 across all samples were kept. The filtered beta value data was subsequently normalized with BMIQ. The final data matrices contain 477,818 probes across 675 samples, and 478,630 probes across 847 samples, for phase-1 and phase-2, respectively.

Zannas

This 450k dataset is derived from whole blood of African-American participants of the Grady Trauma Project [61]. Data was obtained from NCBI GEO website under accession number GSE72680 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72680). The file "GSE72680_beta_values.txt.gz" containing the beta values and detection P-values was downloaded. Probes with P-values < 0.05 across all samples were kept. However, the beta-value matrix of the retained probes still contained NAs and these were imputed with the function impute.knn (k=5) from the *impute* R-package. Beta values were later normalized with BMIQ, resulting in a normalized data matrix containing 453,310 probes across 422 samples.

Flanagan/FBS

The 450k dataset Flanagan et al. is from the Breakthrough Generations Study [62] and was obtained from NCBI GEO website under the accession number GSE61151 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61151). The file "GSE61151_Matrix_raw_signal.txt.gz" containing beta values and detection P-values was downloaded. Only probes with P-values < 0.05 and with no other QC-failures across all samples were kept. The filtered beta value data was subsequently normalized with BMIQ. The 2 duplicates of 4 pairs of samples were averaged, since duplicate pairs exhibited strongest correlations with each other. The final normalized data matrix was defined for 426,430 probes and 184 samples.

Johansson

The 450k dataset from Johansson et al. [63] was obtained from NCBI GEO website under accession number GSE87571 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87571). The file "GSE87571_RAW.tar" containing the IDAT files was downloaded and processed with *minfi* R-package. Probes with P-values < 0.05 across all samples were kept. Filtered data was subsequently normalized with BMIQ, resulting in a normalized data matrix for 475,069 probes across 732 samples.

Luo *et al. Genome Medicine*   (2023) 15:59

Page 6 of 24

TD7k

This EPIC DNAm dataset from TruDiagnostic Inc. was collected between 2020 and 2022. The dataset represents whole blood samples collected from a total of 7719 individuals who had provided blood as part of either a routine check by physicians or by acquiring a kit directly from TruDiagnostic Inc. All individuals have provided consent to use the collected data for this project. Whole blood samples were collected and stored at $-80°C$ prior to DNA processing, which was conducted at the TruDiagnostic Inc. lab facility (Lexington, KY, USA). Five hundred nanograms of DNA was extracted and bisulfite converted using the EZ DNA Methylation kit (Zymo Research) using the manufacturer's instruction. After bisulfite conversion, converted DNA were hybridized to the Illumina HumanMethylation EPIC Beadchip, stained, washed, and imaged with the Illumina iScan SQ instrument to obtain raw image intensities. Raw data was processed using the *minfi* pipeline. Low-quality samples were identified using the ENmix qcfilter() function. Probes with *P*-values < 0.05 across all samples were identified and kept, with low-quality probesets removed. A combinatorial normalization processing using the minfi Funnorm procedure, followed by the RCP method available in ENmix(). The final normalized beta-valued matrix was defined for 864,627 probes across 7719 samples.

Lehne

This 450k DNAm dataset consists of over 2700 peripheral blood samples [64], but we used the already QC-processed and normalized version previously described by Voisin et al. [65] which included a total of 2639 samples.

UCLA

The UCLA dataset (*N* = 178) was collected at Physio-Age LLC and sent to TruDiagnostic Inc. for processing. All processing and data normalization was performed exactly as for the TD7k dataset.

TZH, Hannum, LiuRA, Tsaprouni, MRC1946, and FCE

The TZH (EPIC) [66], Hannum (450k) [67], LiuRA (450k) [5], Tsaprouni (450k) [68], MRC1946 (450k) [18], and FCE (EPIC) datasets [69] were downloaded and normalized as described by us previously [18, 66, 70].

Mass General Biobank data (MGB)

A total of 4386 whole blood samples were retrieved from the MGB Biobank to study associations of immune-cell fractions with health outcomes (not part of the meta-analysis). MGB-derived DNA samples were processed by TruDiagnostic Inc. lab facility (Lexington, KY, USA) in conjunction with the TD7k samples, as described for the TD7k cohort above. The same processing and normalization steps as used in the TD7k cohort were used here. The final processed dataset was used for subsequent association analyses with health outcomes.

Real DNAm datasets with matched FACS cell counts
UCLA

UCLA Immune Assessment Core performed the analysis of immunosenescent cells for 144 whole blood samples, as described previously [71]. Briefly, Total CD3 + T-cells, CD4 + T-cells, CD8 + T-cells, CD19 + B-cells, and CD56 + /CD16 + NK-cells were enumerated in EDTA whole blood with the BD Multitest 6-color TBNK reagent and BD Trucount tubes following the manufacturer's instructions, acquired on a BD FACSCanto II and analyzed with the BD FACSCanto Software. CD8 + T-cell sub setting was performed by staining 50 μl of EDTA whole blood with CD3 FITC, CD8 PerCP, CD28 PE, and CD95 APC (BD) for 10 min, followed by BD FACS Lysing used according to the manufacturer's instructions. At least 10,000 lymphocyte events per sample were acquired and analyzed using DIVA 8.0 software on BD FACSCanto II.

Koestler

We used the Illumina 450k dataset from Koestler et al. [26] consisting of 6 whole blood (WB) with matched flow-cytometric cell counts. This dataset is available from GEO under accession number GSE77797. DNAm data was normalized and processed as previously described [21].

Estimation of cell-type fractions

In all cases, given the 12 cell-type DNAm reference matrix for either the Illumina 850k or 450k dataset, we estimated corresponding cell-type fractions using the EpiDISH Bioconductor R-package [21, 72]. Specifically, we ran the *epidish* function with "RPC" as the method and maxit = 500.

Meta-analyses

In each cohort, associations between CTFs and phenotypes were assessed using multivariate linear regression. Covariates generally included age, sex, smoking status, and batch if evidence for batch effects was present and if batch information was available. In general, however, we note that cell-type fractions are relatively robust to batch effects. For specific cohorts where additional covariates were available, multivariate regression models with these additional covariates were also performed. Smoking status was generally treated as ordinal

with $0 =$ never-smoker, $1 =$ ex-smoker and $2 =$ current smoker. For each regression, we extracted the corresponding effect size, standard error, Student's t test, and *P*-value. Meta-analysis was first performed using the directional Stouffer method: for each study, we transformed the *P*-value into a normal quantile z-value, using the sign of the statistic to assign the sign of the z-value. We then averaged the z-values over studies using the formula $Z = \frac{1}{\sqrt{K}} \sum_{s=1}^{K} z_s$ where K is the number of studies. From Stouffer's z-value an overall *P*-value can then be derived using a standard normal distribution. In addition, we also performed a fixed effect inverse variance meta-analysis [73] using the *metagen* function implemented in the *meta* R-package [74].

Heteroscedasticity between immune-cell fractions and age was assessed using the Breusch-Pagan test, as implemented with the *bptest* function in the *lmtest* R-package [75]. A robust linear estimator was used as implemented in the *rlm* function of the MASS R-package.

scRNA-Seq dataset of peripheral blood mononuclear cells (PBMCs) from 900 donors
scRNA-Seq data from Yazar et al. [76] was downloaded from https://cellxgene.cziscience.com/collections/dde06 e0f-ab3b-46be-96a2-a8082383c4a1. Specifically, we downloaded a Seurat object including the cell-by-gene expression matrix as well as associated metadata containing cell type annotation, matched sample IDs, age, sex and UMAP coordinates for 1,248,980 PBMCs from a total of 981 donors. Of the 31 cell types, some were merged as required for comparison with DNAm: CD4+TCM and CD4+TEM were regarded as CD4+T memory cells; CD8+TCM and CD8+TEM as CD8+T memory cells; cells labeled as "NK_CD56bright," "NK Proliferating," and "NK" were regarded as NK-cells; CD14+monocytes and CD16+monocytes were treated as monocytes; cell types not included in our DNAm reference were regarded as "other." After this merging, we were left with 9 cell types (not including "other"): approximately 320k CD4+T memory cells, 259k naïve CD4+T-cells, 177k CD8+T memory cells, 52k naïve CD8+T-cells, 171k NK, 65k naïve B-cells, 30k memory B-cells, 51k monocytes, and 26k T-regulatory cells. Cell type proportions and differential abundance of these in relation to age and sex were calculated with the *propeller* method [77] from *speckle* R-package.

scRNA-Seq dataset of mild and severe Covid-19 cases
scRNA-Seq expression data of bronchoalveolar lavage fluid (BALF) from Liao et al. [78] encompassing 3 COVID-19 mild samples, 6 COVID-19 severe samples, and 4 healthy controls was obtained from https://cells.ucsc.edu/covid19-balf/exprMatrix.tsv.gz. We also downloaded the corresponding metadata containing

annotation for cell types and samples from https://cells.ucsc.edu/covid19-balf/meta.tsv. QC for scRNA-Seq data has been done by Liao et al. by only keeping cells with gene number between 200 and 6000, UMI count > 1000, and mitochondrial gene percentage < 0.1. Cells were annotated after batch effect removal with FindIntegrationAnchors and IntegrateData functions and clustering with FindNeighbors and FindClusters functions from Seurat package [79]. There were clusters mapping to 49417 macrophages, 7716 T-cells, 220 B-cells, 1607 neutrophils, 3531 epithelial cells, 70 mast cells, 978 mDCs, 1081 NKs, 152 pDCs, and 1041 plasma cells, each cluster annotated with signature genes. We used the scRNA-Seq data defined over 23,916 genes and T-cells, B-cells, neutrophils, and NKs from 11 samples (3 healthy, 3 mild, 5 severe) for the following analysis, normalizing the data with NormalizeData(normalization.method = "LogNormalize," scale.factor = 10^4) from Seurat package. Note that among the original 13 samples, one healthy sample (labeled as "HC2" in metadata provided by Liao et al.) and one severe sample (labeled as "S3" in metadata provided by Liao et al.) were removed due to small numbers of T-cells in these two samples (< 70 cells). T-cells were classified as "naïve T-cells" if the expression values of LEF1 are non-zero (948 naïve cells), otherwise classified as "memory/effector T-cells" (6653 memory/effector cells). We used the function *propeller(robust = FALSE, trend = FALSE, transform = "asin")* from speckle R-package to calculate for each cell-type their proportions, to perform a variance stabilizing transformation on the proportions and to determine whether the differential abundance is statistically significant between non-severe and severe samples.

Health outcome cox-regression analysis in the MGBB cohort
We queried the demographic information (i.e., date of birth, sex and ethnicity), health history (i.e., smoking status, alcohol consumption and BMI), and clinical records (i.e., patient diagnosis) of 4386 human subjects from Mass General Brigham (MGB) Biobank [80] and The Research Patient Data Repository (RPDR) databases [81]. Age at the time of sample collection was then calculated accordingly. The vital status (i.e., living/deceased) and date of death were also obtained from MGB Biobank. However, 147 subjects who were recorded as deceased had missing date of death, and they were excluded from the survival analysis of all-cause mortality. We identified other diseases, including type 2 diabetes, chronic obstructive pulmonary disease (COPD), cardiovascular disease (CVD), cancer, and depression, by using relevant ICD-9/10 diagnosis codes (referred to supplement codebook). We defined

Luo *et al. Genome Medicine* (2023) 15:59

Page 8 of 24

an incident case as the first diagnosis of a specific health outcome that occurred on the patient's medical record after the sample collection date. Subjects with a diagnosis code of diseases prior to sample collection were excluded from the survival analysis of that particular disease. We imputed missing data for smoking status, alcohol consumption, and BMI by utilizing the longitudinal records of these variables. Specifically, we used the record closest to the sample collection date for smoking status and alcohol consumption. For BMI, we imputed the missingness as the median value of all BMI records within 6 months around the collection date to balance off the measurement error and temporal variation. Despite imputation, 605 subjects still had missing BMI data and were excluded from the survival analysis when further adjusting for additional risk factors, including BMI. We estimated the hazard ratio of each immune cell type against the health outcomes using Cox-proportional hazard regression models with *coxph* function in *survival* R-package. The models were adjusted for age, sex, ethnicity, and baseline comorbidities (which included cancer, CVD, COPD, depression, and T2D), and separately again adjusting for age, sex, ethnicity, smoking status, alcohol consumption, BMI, and the same baseline comorbidities.

Lasso penalized Cox-regression model predictor of all-cause mortality

Using the same MGBB cohort, we built predictors of all-cause mortality using all 12 immune-cell fractions, and separately again using in addition also age, ethnicity, sex, BMI, smoking status, alcohol consumption, and all underlying comorbidities. We used a penalized (lasso penalty) Cox-proportional hazard regression model as implemented in the *glmnet* R-package. Briefly, we divided the dataset up into a 70% training (3591 samples + 302 events) and 30% test (1110 samples + 122 events) set. On the 70% training set, we applied an internal tenfold cross-validation procedure [82], to obtain a risk score for each left-out bag in turn and for each choice of penalty

parameter value. The risk scores were then combined across all left-out bags, and the association with all-cause mortality assessed using the C-index. This yielded a curve of how the C-index varies as a function of penalty parameter. In the case of the model that only includes immune-cell fractions, we obtained one clear optimal model. In the case of the model that included all factors, we considered the top 2 model with overlapping 95%C CIs. These models were then tested on the 30% test-test. In all cases, we recorded the hazard ratio, C-index, and their 95% confidence intervals. *P*-values of association between the risk scores and all-cause mortality were derived from the one-tailed Chi-square test (1 degree of freedom) as applied to the Cox-score statistic.

Results

Construction and validation of a 12 immune-cell-type DNAm reference matrix

We constructed a novel DNAm reference matrix encompassing 12 blood cell subtypes (monocytes, neutrophils, eosinophils, basophils, naïve and memory CD4 + T-cells, naïve and memory CD8 + T-cells, naïve and memory B-cells, natural killer (NK) and T-regulatory (Tregs) cells) using the EPIC DNAm profiles of FACS-sorted cells from Salas et al. [28]. Since cell-type-specific marker genes display a very strong preference for unmethylated promoters and enhancers in the corresponding cell types [24, 29], we decided to construct a DNAm reference matrix focusing on CpGs specifically unmethylated in each cell type (Methods). For each of the 12 cell types, we selected CpGs significantly hypomethylated (FDR < 0.001) in the given cell-type relative to all the rest, subsequently ranking them according to the difference in average DNAm, so as to ensure maximum separability (Fig. 1a, "Methods"). We verified that by selecting the 50 top-ranked CpGs for each cell type (i.e., a total of 600 CpGs), that these displayed very significant hypomethylation and relatively big differences in average DNAm, as required (Fig. 1a, Additional File 1: table.S1). About half of the 600 CpGs mapped to gene bodies, while the other

(See figure on next page.)

Fig. 1 Construction and validation of the 12 blood cell-type hypoDNAm reference matrix. **a** Left panel: Example of a CpG's DNAm profile in the DNAm reference matrix, marking T-regulatory (Treg) cells. The *y*-axis labels the cell types, *x*-axis labels the DNAm value, and the number of samples of each cell type (i.e., in each boxplot) is shown on the *y*-axis. Right panel: The DNAm reference matrix for 12 blood cell subtypes encompassing 600 CpGs (i.e., 50 markers per cell type). **b** Scatterplots of true fractions vs estimated fractions for 10 blood cell subtypes using the EPIC DNAm data from 10 artificial mixtures where the underlying mixing proportions were known. For each estimated cell type, we display the *R*-value (Pearson correlation coefficient) and root mean square error (RMSE). **c** Heatmap displays the estimated fractions of cell-sorted samples for each of the 12 immune-cell subtypes in our Illumina 450k/850k DNAm reference matrices, 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 study from which the sorted sample derives from is indicated by the color bar below the heatmap. The technology used to generate the DNAm data of the sorted sample is also indicated. For the 450k and WGBS samples, we used the 450k and 850k DNAm reference matrices, respectively, to obtain the fractions. The estimated fractions in the heatmap are median values taken over biological replicates of the cell-sorted samples, with the number of corresponding biological replicate samples indicated at the bottom
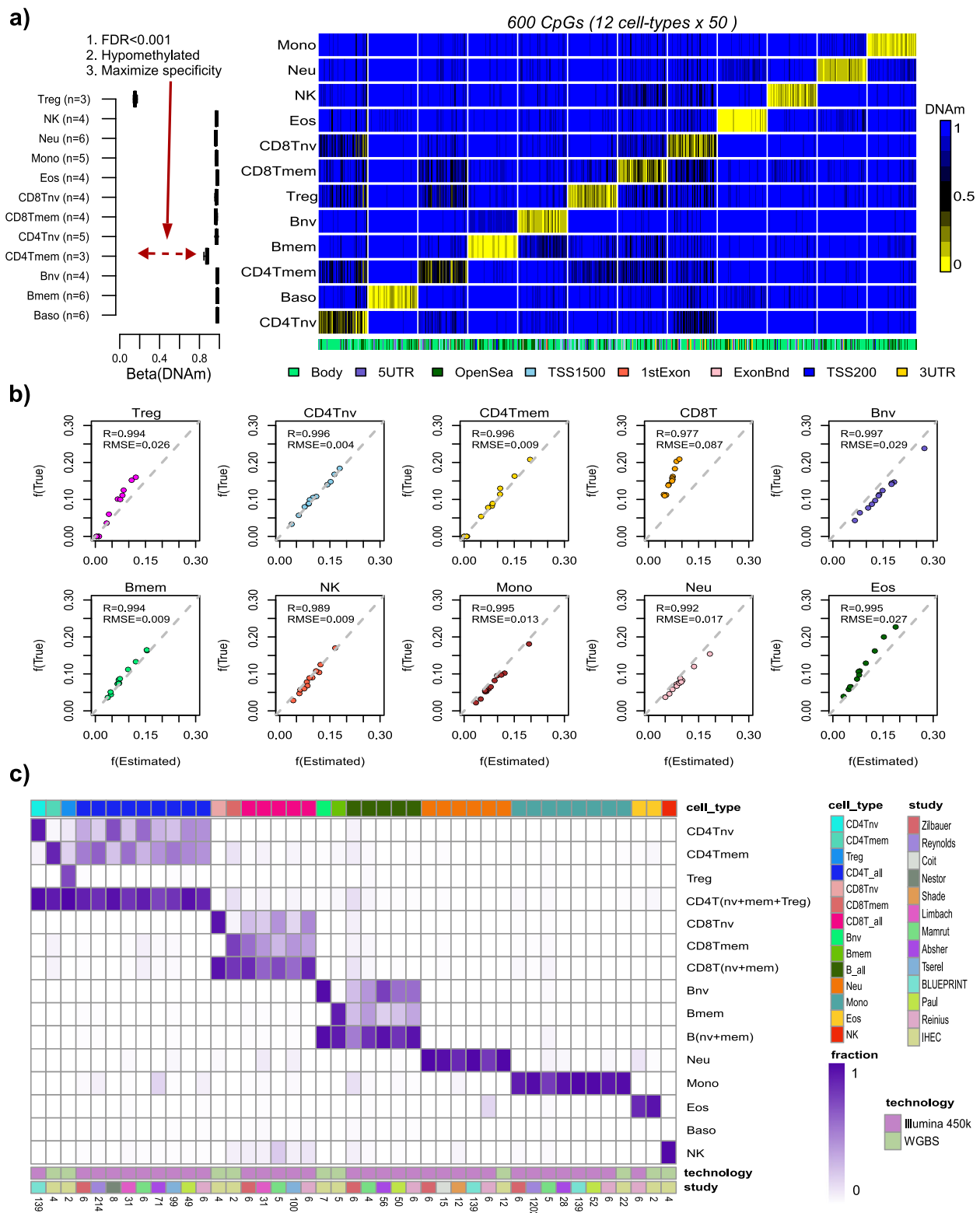
Fig. 1 (See legend on previous page.)

Luo *et al. Genome Medicine*   (2023) 15:59

Page 10 of 24

half mapped predominantly to inter-genic regions and shores/shelves upstream of the TSS (Fig. 1a).

To validate the 12 blood cell-subtype DNAm reference matrix, we applied it in conjunction with the robust partial correlation (RPC) framework implemented in EpiDISH [21] to estimate cell-type fractions in 12 artificial mixtures where the proportions of 10 underlying blood cell subtypes used in these mixtures are known [28]. Pearson correlations and root mean square error (RMSE) were excellent (R ~ 0.99, RMSE < 0.05) for all tested cell types, except for CD8 + T-cells which displayed marginally worse values (R ~ 0.98, RMSE = 0.09) (Fig. 1b). To further validate it, we compared estimated cell-type fractions with matched flow-cytometric counts in two independent datasets, one encompassing 6 blood samples with matched counts for 7 blood cell subtypes [25], and another encompassing 144 peripheral blood samples with matched counts for 3 types of lymphocytes (Methods). For both datasets, Pearson correlation R-values and RMSE were reasonably good (R > 0.84 & RMSE < 0.05 and R > 0.7 & RMSE < 0.1, respectively), further attesting to the quality of our DNAm reference matrix (Additional File 2: fig.S1a-b). For completeness, we repeated the construction of a 12 cell-type DNAm reference matrix, but this time restricting to Illumina 450 k probes, resulting in a separate 600 CpG × 12 cell-type DNAm reference matrix (Additional File 1: table S2). We successfully validated this DNAm reference matrix in artificial blood mixtures and in blood samples with matched flow-cytometric counts (Additional File 2: fig.S2). Of note, estimated cell-type fractions for 7 pooled immune cell types, as derived from each of the two 12 immune-cell-type DNAm reference matrices, displayed good agreement with those derived from our previous 7 immune-cell-type DNAm reference [21], as assessed in the two largest cohorts (TD7k & Lehne) (Additional File 2: fig.S3). In a few cases where correlations were weaker (e.g., R = 0.55 for eosinophil fraction in Lehne et al., see Additional File 2: fig.S3), this was driven by a significant number of unrealistic zero eosinophil fractions as derived with the lower resolution 7 cell-type DNAm reference matrix. This attests to the improved inference possible with a higher-resolution DNAm reference matrix. Further supporting this, both 12 immune-cell-type DNAm reference matrices displayed stronger correlations than the 7 cell-type one, with known experimental cell fractions of artificially generated mixtures, as well as with flow-cytometric counts of whole blood samples (Additional File 2: fig. S4). Finally, we further validated both 850k and 450k versions of the DNAm reference matrix in a large collection of sorted immune-cell subsets, including whole-genome bisulfite sequencing (WGBS) samples from IHEC [83] (Fig. 1c). Of note, our 850k DNAm reference matrix achieved remarkably high accuracy (mean R-value > 0.95) on in silico mixtures generated from these WGBS immune-cell-sorted samples (SI fig.S1c, Methods), thus demonstrating that our DNAm reference matrix built with Illumina DNAm data is applicable to WGBS data.

A meta-analysis of immune-cell fractions reveals novel associations with age

We next applied the 12 immune-cell type DNAm reference matrix to perform a large meta-analysis of immune-cell fractions with common phenotypes. This meta-analysis serves two purposes. First, because large DNAm datasets with matched flow-cytometric counts for as many as 12 immune cell types are not available, it is paramount to seek additional means to validate our high-resolution DNAm reference matrix on real data. By estimating immune-cell fractions in a large number of independent cohorts and correlating these to specific phenotypes, we can ascertain the quality of our novel DNAm reference matrix. For instance, our DNAm reference matrix should be able to correctly capture a well-known age-associated immunosenescence signature characterized by a decreased naïve to mature T-cell fraction ratio [6, 84–89]. Second, a meta-analysis can reveal subtle, novel, and highly statistically significant associations, not evident from individual studies. To perform the meta-analysis, we estimated fractions for all 12 immune-cell subtypes in 22 independent whole blood cohorts, encompassing 23,053 samples and two versions of the Illumina DNAm array (EPIC & 450k) (Additional File 1: tables S3-S4). In each cohort, we correlated these fractions to common phenotypes using multivariate linear regression models that adjust for study-specific confounders ("Methods," Additional File 1: table S5). We used the directional Stouffer method to derive an overall *z*-statistic and *P*-value of association across all studies with available phenotype information ("Methods").

We first considered the case of age. The chronological age distribution was reasonably wide (age range > 30 years) for all 22 studies, except for one where all individuals were of the same age and which was henceforth excluded from age-association analyses (Additional File 2: fig.S5, Additional File 1: table S3). Validating our DNAm reference matrix, we observed a strong consistent reduction in the naïve CD8 + T-cell population with age across all 21 studies (Fig. 2a, Stouffer $Z = -53$, $P < 10^{-200}$). For CD4T + cells, the reduction of the naïve subset was also evident in 17 out of 21 studies (Fig. 2a, Stouffer $Z = -25$, $P < 10^{-100}$). Correspondingly, there was a trend for the memory

T-cell subsets to increase with age (Fig. 2a, Stouffer $Z=14$, $P<10^{-40}$ for CD4Tmem and $Z=9$, $P<10^{-20}$ for CD8Tmem). A fixed effect inverse variance meta-analysis model showed that effect sizes were small, typically involving at most only a few percentage points over a 50-year interval (Fig. 2b,c, "Methods," Additional File 2: fig.S6). A formal test for heterogeneity of effect size (I^2) revealed substantial heterogeneity between cohorts (Fig. 2b), which is unsurprising given the diverse nature of the cohorts included in the meta-analysis. Overall, we observed an excellent agreement between the directional Stouffer and fixed effect meta-analysis model (Fig. 2d). Although deviations from homoscedasticity were evident (Additional File 2: fig.S7), for instance, the naïve CD8+T-cell fraction displayed decreased variance with age in most of the cohorts examined, t-statistics were effectively unchanged when rerunning the multivariate linear regressions with a robust Huber M-estimator (Additional File 2: fig.S8).

 A meta-analysis over many datasets can also reveal novel associations or strengthen previous preliminary findings. For instance, our meta-analysis revealed a clear trend for basophil and NK-cell fractions to increase with age (Fig. 2a, Stouffer $Z=11$, $P<10^{-20}$ for NK and $Z=10$, $P<10^{-20}$ for basophils), strengthening preliminary findings from others [90–93]. To further validate the observed associations for lymphocytes and monocytes, we collated a large scRNA-Seq dataset of over 1.27 million peripheral blood mononuclear cells (PBMCs) from over 900 donors spanning a wide age range (Fig. 2e, "Methods") [76], and used the propeller DA-testing method [77] to derive statistics of association of relative immune-cell fractions with age. This revealed an excellent agreement between the predictions from DNAm data and scRNA sequencing (Fig. 2f,g). Our meta-analysis also revealed that the eosinophil fraction did not change significantly with age (Fig. 2a, Stouffer $P=0.03$), the marginal significance being driven entirely by one study (Song et al. [53]). Of note, this study had profiled DNAm in childhood cancer

survivors, and the observed increased eosinophil count in Song et al. was independent of cancer treatment type (Additional File 2: fig.S9). Our eosinophil meta-analysis result validates a major eosinophil count study in over 11,000 subjects, which concluded that eosinophil counts do not change with age beyond the age of puberty [94]. Consistent with this, all cohorts analyzed here did not include samples pre-puberty except for the large TD7k cohort which however only included 3 samples younger than 15 years (Additional File 2: fig.S5b).

Decrease of naïve CD8+T-cell fraction with age is strongest before the age of 40

A large collection of DNAm datasets in blood can also help address the question if age-associated changes in immune-cell fractions are a linear process. We focused on the naïve T-cell fractions because these displayed the most consistent and significant changes with age. To address the question, we merged the estimated fractions of samples in the age range 20 to 85 from all cohorts where the given cell-type fraction was significantly associated with age (with same directionality), excluding the Song et al. cohort since this was a PBMC dataset displaying significantly different levels of cell-type fractions (Additional File 2: fig.S10). In the case of the naïve CD8+T-cell fraction, we observed that the decrease was more accentuated for younger individuals in the age range 25–40, with a shift in the gradient at approximately age 40, followed by a more moderate decrease in the age range 40 to 60, with no clear further decrease beyond the age of 60 (Fig. 3a). We verified that this pattern was present in both sexes separately (Fig. 3a) and in each of the largest cohorts with sufficient and balanced representation across age groups, demonstrating that the more rapid decrease before age 40 is not a technical artefact of the merging procedure (Additional File 2: fig.S11). In contrast, the naïve CD4+T-cell fraction displayed a more constant rate of decrease throughout life (Fig. 3b).

(See figure on next page.)

Fig. 2 Association and validation of immune-cell-type fractions with age. **a** Heatmap of associations between blood cell-type fractions and age in each of 21 cohorts. Cohort sizes are shown above heatmap. Colors indicate both directionality of change and statistical significance, as indicated, with the *P*-values derived from a multivariate regression that included sex, smoking status, BMI, and other confounders as described in "Methods." Meta(Stouffer) indicates the directional Stouffer meta-analysis *z*-statistic and statistical significance. Associations marked with an * are significant after Bonferroni correction at 0.05 level (i.e., $P<0.05/11$). **b** Barplot displaying the effect size estimates and 95% confidence intervals from a fixed effect inverse variance meta-analysis model. Estimates have been multiplied by 50 to reflect the percentage change over a 50-year period. Color bar to the right labels the I^2 values of heteroscedasticity. **c** Smoothed scatterplots displaying the naïve CD8+ and CD4+T-cell fractions with age in the TZH cohort. Black dashed line and *P*-value is from a linear regression. **d** Scatterplot of *z*-statistics from the directional Stouffer method, against the corresponding statistics from the fixed effect meta-analysis model. R-value of agreement is shown. Gray dashed lines indicate the line $P=0.05$. **e** UMAP plots of the scRNA-Seq data of Yazar et al. profiling > 1 million PBMCs from over 900 donors of different ages. **f** Scatterplot of Stouffer *z*-statistics for each cell type derived from DNAm data, against the corresponding *t*-statistic from Propeller method based on the scRNA-Seq data. Linear regression *P*-value of agreement is given. **g** Examples of immune-cell fractions displaying significant associations with chronological age as inferred from scRNA-Seq data. *P*-values derive from Propeller
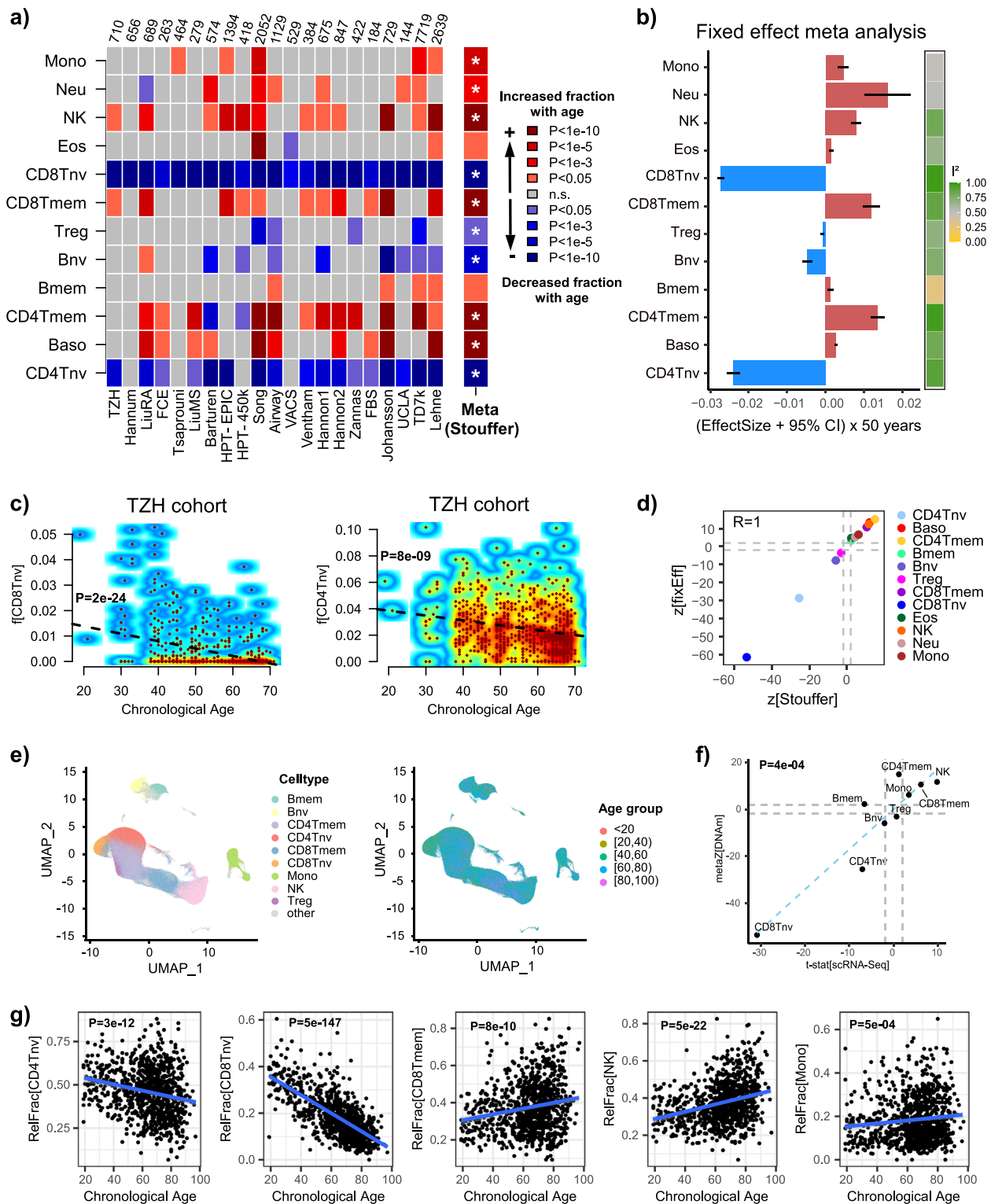
Luo *et al. Genome Medicine*   (2023) 15:59

Page 12 of 24

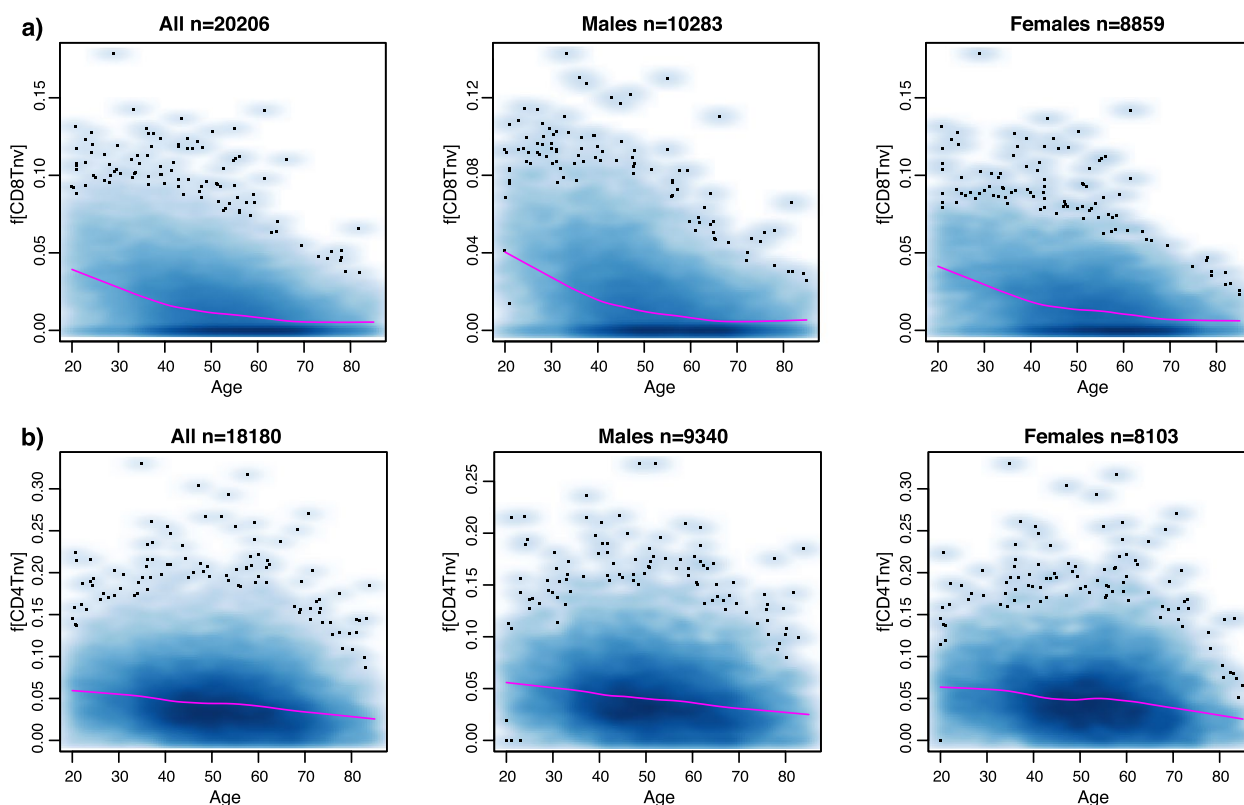**Fig. 2** (See legend on previous page.)

Fig. 3 Non-linear rate of change of immune-cell-type fractions with age. **a** Left panel: Smoothed density scatterplot of the estimated naïve CD8 + T-cell fraction against chronological age for all samples aged between 20 and 85 from 20 cohorts, with the loess regression curve (span = 0.3) displayed in magenta. Excluded cohorts are MRC1946, which only contains samples of the same age (age = 53) and Song, which profiled PBMCs and not whole blood. Middle and Right panels: as left but plotting males and females separately. **b** As **a** but for the estimated naïve CD4 + T-cell fraction. Only samples aged between 20 and 85 from 16 cohorts where the association of naïve CD4 + fraction with age was significant (multivariate linear regression $P < 0.05$) and with same directionality were used. Excluded cohorts are MRC1946, Hannum, Tsaprouni, HPT-450 k, Song, and VACS

Meta-analysis reveals novel associations of immune-cell counts with sex

A total of 18 studies profiled blood in both males and females (Additional File 2: fig.S12a). A meta-analysis using directional Stouffer revealed increased monocyte, eosinophil, basophil. and NK-cell fractions in males, and increased T-regulatory and naïve T-cell subset fractions in females (Fig. 4a, Stouffer $P < 10^{-10}$ for all cell types). Very similar findings were obtained using a fixed effect meta-analysis model (Fig. 4b,c, Additional File 2: fig.S13). Once again, we were able to validate most of the associations using the large scRNA-Seq dataset of PBMCs from Yazar et al. [76] (Fig. 4d–f). The observed increased eosinophil fraction in males is also consistent with a similar finding from a major eosinophil count study [94]. Of particular importance is the observed increased regulatory and naïve T-cell fractions in women (Fig. 4a), an observation that, surprisingly, has not been noted before, except for a sporadic mention in one recent study by Bergstedt et al. [95]. Of note, these particular sex associations were validated using an orthogonal technology (scRNA-Seq data) (Fig. 4d–f), in clear support of their significance and biological relevance, highlighting a novel insight which could be important for understanding sex-specific differences in cancer and autoimmune disease risk [96]. Interestingly, in the only cohort where the association of naïve CD4 + T-cells with sex was not seen (the Han Chinese TZH cohort, Fig. 4a), we verified that this was due to residual confounding between alcohol consumption and sex (Chinese women consume little alcohol), as indeed the association became significant when adjusting additionally for alcohol consumption (Additional File 1: table S6).

Associations of immune-cell fractions with smoking, obesity, exercise, and stress

Smoking status information was available for 10 studies (Additional File 2: fig.S12b). Compared to age and sex, the number of associations with smoking was much lower (Fig. 5a). The most significant and consistent associations were displayed by the CD4 + T-cell memory fraction, which was significantly higher in smokers in at least
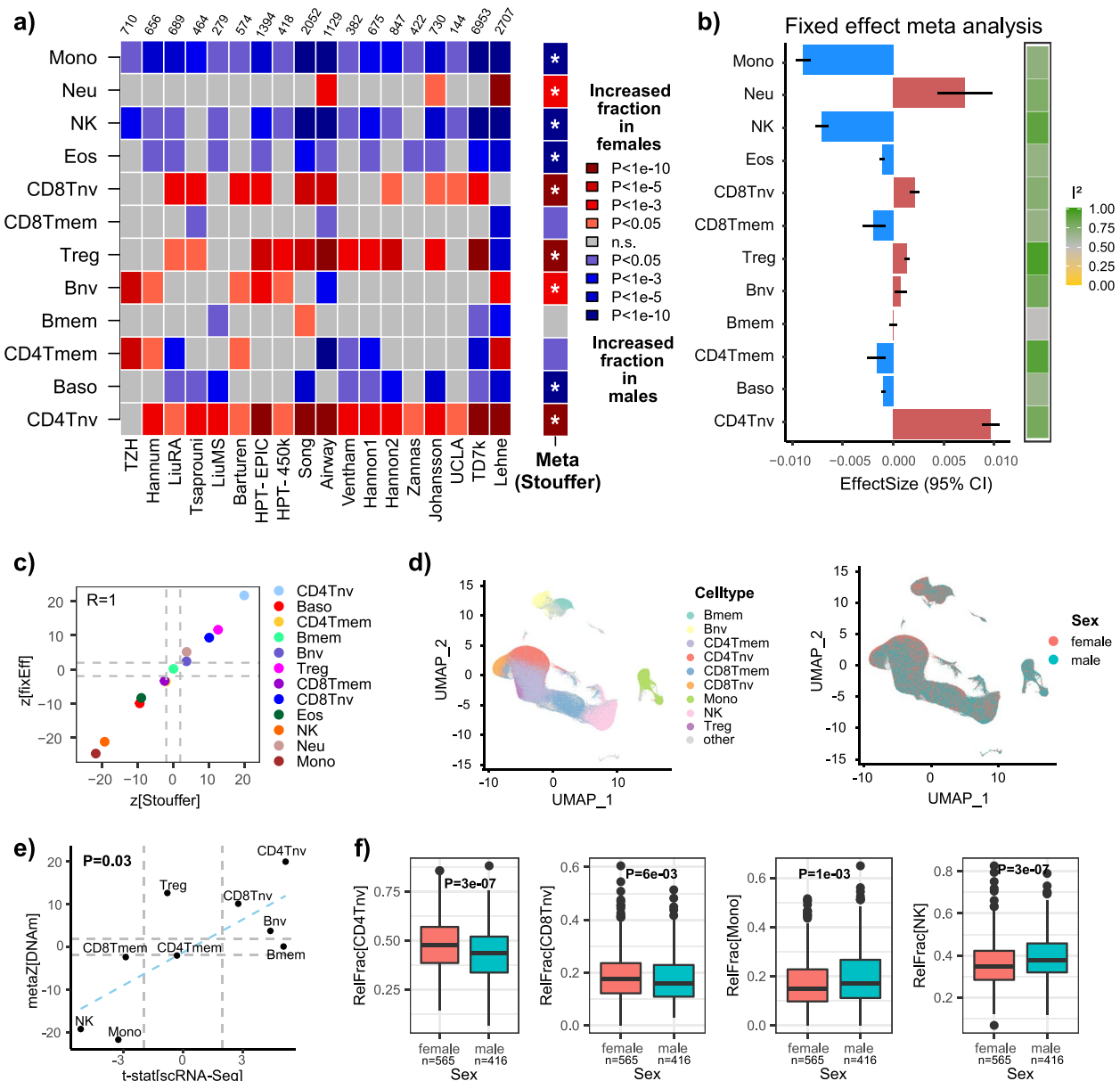
Fig. 4 Association and validation of immune-cell fractions with sex. **a** Heatmap of associations between blood cell-type fractions and sex in each of 18 cohorts with sex information. Cohort sizes are shown above heatmap. Red (blue) tones indicate fractions that increase in females (males). *P*-values derived from a multivariate regression that included age, smoking status, BMI (whenever available), and other study-specific confounders as described in "Methods." Meta(Stouffer) indicates the directional Stouffer meta-analysis *z*-statistic and statistical significance. Associations marked with an * are significant after Bonferroni correction at 0.05 level (i.e., *P* < 0.05/11). **b** Barplot displaying the effect size estimates and 95% confidence intervals from a fixed effect inverse variance meta-analysis model. Color bar to the right labels the *I*² values of heteroscedasticity. **c** Scatterplot of *z*-statistics from the directional Stouffer method, against the corresponding statistics from the fixed effect meta-analysis model. R-value of agreement is shown. Gray dashed lines indicate the line *P* = 0.05. **d** UMAP plots of the scRNA-Seq data of Yazar et al. profiling > 1 million PBMCs from over 900 donors encompassing both sexes as shown. **e** Scatterplot of Stouffer *z*-statistics for each cell type derived from DNAm data, against the corresponding *t*-statistic from Propeller method based on the scRNA-Seq data. Linear regression *P*-value of agreement is given. **f** Examples of immune-cell fractions displaying significant associations with sex as inferred from scRNA-Seq data. *P*-values derive from Propeller
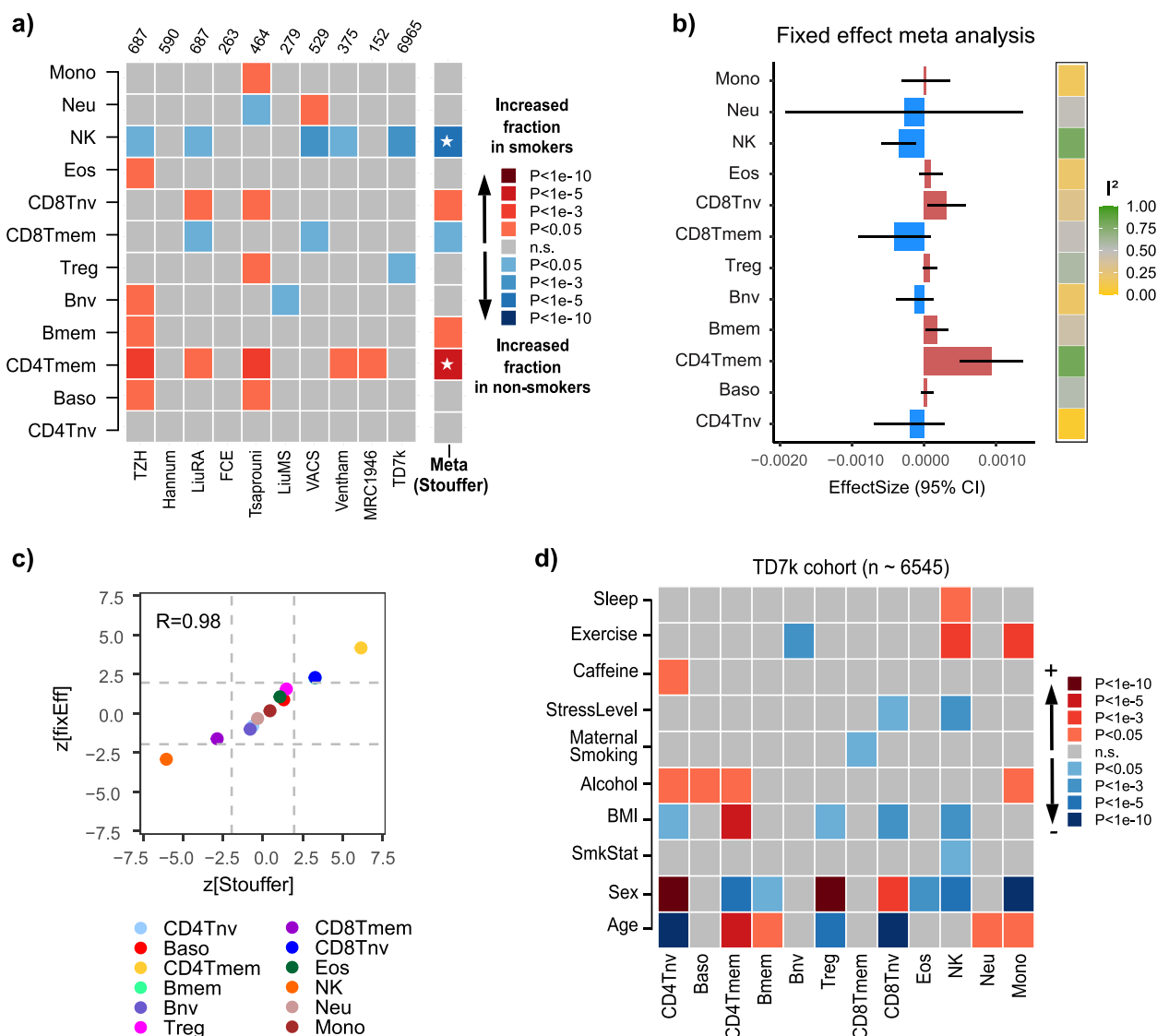
Fig. 5 Association of immune-cell fractions with smoking, BMI, exercise, and stress. **a** Heatmap of associations between blood cell-type fractions and smoking in each of 10 cohorts with smoking status information. Cohort sizes are shown above heatmap. Red (blue) tones indicate fractions that increase in smokers (never-smokers). *P*-values derived from a multivariate regression that included age, sex, and BMI (whenever available) and study-specific confounders as described in "Methods." Meta(Stouffer) indicates directional Stouffer *z*-statistic and significance. Associations marked with a * are significant after Bonferroni correction at 0.05 level (i.e., *P* < 0.05/11). **b** Barplot displaying the effect size estimates and 95% confidence intervals from a fixed effect inverse variance meta-analysis model. Color bar to the right labels the I^2 values of heteroscedasticity. **c** Scatterplot of *z*-statistics from the directional Stouffer method, against the corresponding statistics from the fixed effect meta-analysis model. *R*-value of agreement is shown. Gray dashed lines indicate the line *P* = 0.05. **d** Heatmap of multivariate associations of immune-cell fractions with epidemiological factors in the TD7k cohort, encompassing over 6545 samples. Red tones indicate fractions that increase with increasing values of the epidemiological factors, blue tones indicate decreases. *P*-values were derived from a multivariate linear regression including all phenotypic factors as shown

6 of the 10 studies (Fig. 5a, Stouffer $Z = 6$, $P = 4 \times 10^{-9}$). In 5 of 10 studies, we also observed a consistent decrease of the NK-cell fraction in smokers (Fig. 5a, Stouffer $Z = -5$, $P = 2 \times 10^{-6}$). These two associations were also observed under a fixed effect meta-analysis model (Fig. 5b,c, Additional File 2: fig.S14). Although smoking information was

not available for the scRNA-Seq study of Yazar et al., the findings are nevertheless consistent with previous blood cell-count studies reporting increased CD4 + T-cell memory and decreased NK counts in smokers [90, 97–99].

BMI information was available for 5 studies (Additional File 2: fig.S12c), but the only highly significant

associations with BMI were seen in the largest of these cohorts (TD7k), encompassing 6545 samples (Additional File 2: fig.S15). For the TD7k cohort, additional covariate information was available for most of these 6545 samples; hence for this cohort, we performed a more extensive multivariate linear regression analysis including not only age, sex, and smoking status as covariates, but also sleep, physical exercise, caffeine and alcohol consumption, and maternal smoking exposure. Interestingly, as with smoking, we observed an increased CD4 + T-cell memory fraction in obese individuals (Fig. 5d, linear regression $P = 1 \times 10^{-8}$), which was also seen in the second largest cohort with BMI information (TZH cohort) (Additional File 2: fig.S15). Of note, in the TD7k cohort, naïve T-cell and NK-cell fractions were reduced in obese individuals independently of all other covariates (Fig. 5d, linear regression $P = 0.0004$ (CD8Tnv), $P = 0.016$ (CD4Tnv), and $P = 0.0004$ (NK)). The multivariate analysis in TD7k also revealed a significant increase of NK (linear regression $P = 6 \times 10^{-5}$) and monocyte fractions (linear regression $P = 0.0008$) in individuals undergoing frequent exercise, while the naïve B-cell fraction decreased (linear regression $P = 0.0001$) (Fig. 5d). The NK fraction also increased in individuals sleeping longer hours although this association was more marginal (linear regression $P = 0.003$, Fig. 5d). In contrast, the NK fraction decreased significantly in individuals reporting higher stress levels (linear regression $P = 0.0006$, Fig. 5d). Although associations with alcohol consumption were only of marginal significance, it is worth noting that the specific increases of CD4T cells (both naïve and memory subsets) and basophils in frequent drinkers were replicated at a similar significance level in the TZH cohort (Additional File 2: fig. S16). Overall, these patterns reveal associations of specific cell types (e.g., memory and naïve T-cells, NK) with multiple independent factors (smoking, obesity, stress, exercise, and spirit drinking).

Associations of immune-cell fractions with health outcomes

To explore associations of immune-cell fractions with health outcomes, we focused on a cohort derived from the Mass General Brigham Biobank (MGBB) (Methods). This cohort has extensively annotated epidemiological and prospective health outcome information, including all-cause mortality, type-2 diabetes (T2D), cancer, chronic obstructive pulmonary disease (COPD), stroke, cardiovascular disease (CVD), and depression (Methods) for 4386 subjects. The demographics of this cohort and their underlying comorbidities at baseline (blood sample draw) is shown in Additional File 1: table S7. In the first instance, we used Cox-proportional hazard regressions to evaluate associations between each of the 12

immune-cell fractions with each of the various outcomes, adjusting for intrinsic factors that included age, sex, and ethnicity (Methods). In the case of all-cause mortality, we also adjusted for co-morbidity at baseline. This was done by including separate covariates for depression, COPD, type 2 diabetes, CVD, and cancer defined at baseline or before sample draw (Methods). We observed many associations (Fig. 6a, Additional File 1: tables S8-S9) that remained significant upon further adjustment for additional disease risk factor exposures including smoking, obesity, and alcohol consumption (Fig. 6b, Additional File 1: tables S10-S11). Notably, naïve CD4 + T-cell, naïve B-cell, and NK-cell fractions were all associated with a reduced risk of all-cause mortality, even after adjustment for all major disease risk factors and baseline co-morbidity. Interestingly, while the naïve CD4 + T-cell fraction also displayed negative associations with many health outcomes, notably with COPD and T2D, the memory CD4 + T-cell fraction was only negatively associated with all-cause mortality. Some of the other associations were strongly dependent on the specific health outcome. For instance, an increased memory B-cell fraction was specifically associated with an increased risk of cancer but a reduced risk for T2D, while no associations were observed for the other outcomes. These results highlight the potential biological importance of immune-cell fractions in disease risk.

To understand how well immune-cell fractions can predict all-cause mortality, we computed corresponding C-index values, revealing that individual fractions can only modestly predict outcome (Additional File 1: table. S12). For instance, the naïve CD4T-cell fraction displayed a C-index of 0.65 (95%CI 0.62–0.68). Next, we asked how well a multivariate model that only includes immune-cell fractions would fare. To this end, we split the dataset into a 70% training (2591 samples, 302 events) and 30% test set (1110 samples, 122 events) and trained a lasso Cox-proportional hazards regression model on the training set using a tenfold internal cross-validation strategy (Methods). Using the C-index as the performance metric, we identified an optimal model in the training set which achieved a C-index value of 0.69 (95% CI 0.67–0.72) in the blinded test set (Additional File 1: table.S12). Upon further inclusion of age, sex, ethnicity, smoking, BMI, alcohol consumption, and all co-morbid conditions, we identified two optimal models with overlapping 95% confidence intervals: one included only age and naïve CD4 + T-cell fraction as covariates, while another included additional variables, notably sex and co-morbidity for all conditions (Additional File 1: table S13). These models achieved hazard ratio (HR) and C-index values of HR = 2.96 (95%CI 2.54–3.44) & C = 0.75 (95%CI 0.72–0.77) and HR = 3.86 (95%CI 3.32–4.48) & C = 0.81
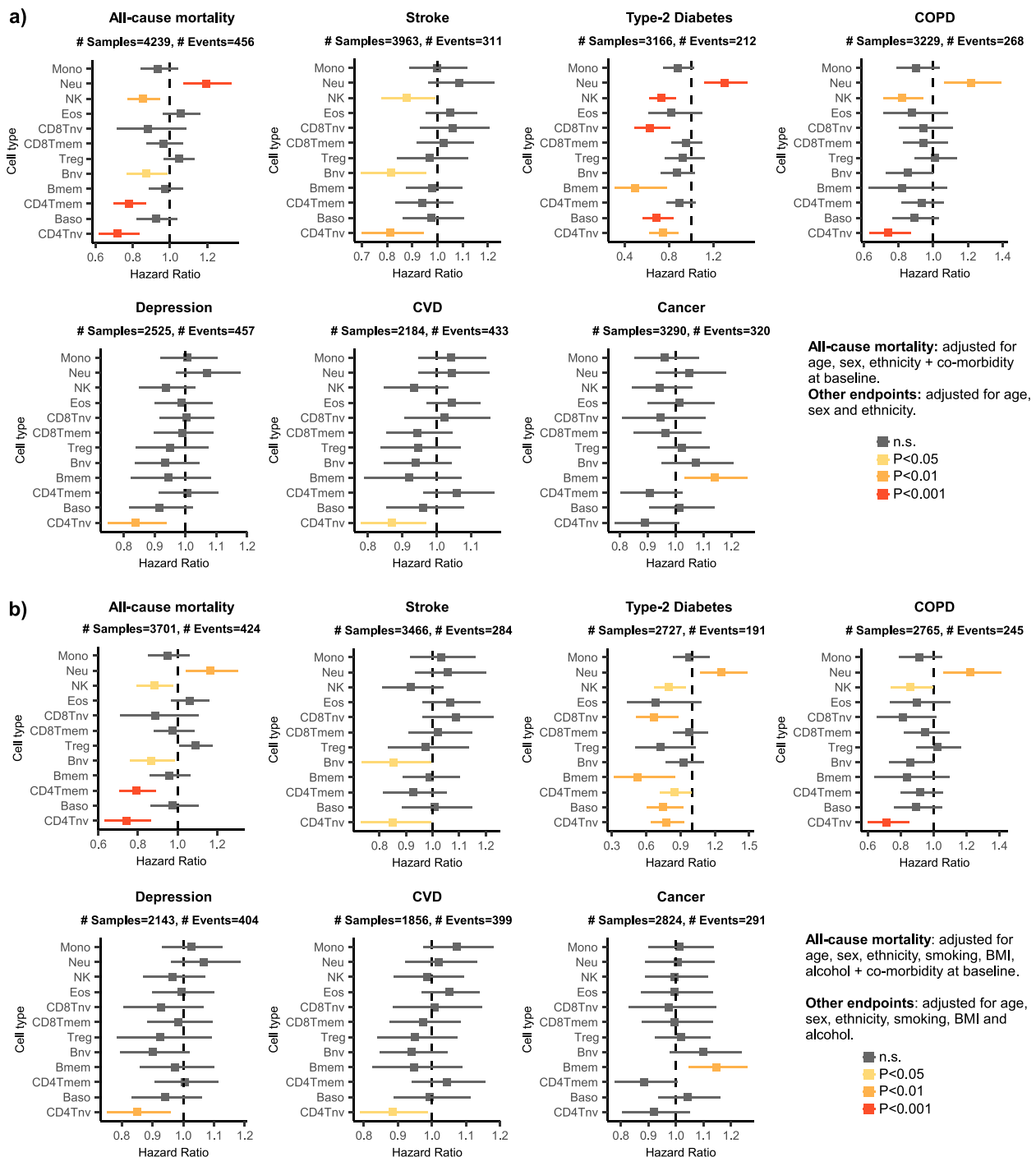
Fig. 6 Association of immune-cell fractions with health outcomes in MGB cohort. **a** Forest plots of association between immune-cell fractions with various health outcomes, as shown. In each panel, the *x*-axis labels the hazard ratio (HR) as evaluated using a Cox-regression model that included age, sex, and ethnicity as covariates. In the case of all-cause-mortality, we also adjusted for co-morbidity at baseline. *P*-values derive from a two-tailed Wald-test. Vertical dashed line indicates the line HR = 1. For each HR datapoint, we display the 95% confidence interval. For each health outcome, we display the total number of samples and events. **b** As **a**, but adjusting in the Cox-regression model also for smoking, BMI, and alcohol consumption

Luo *et al. Genome Medicine*   (2023) 15:59

Page 18 of 24

(0.79–0.83), respectively (Additional File 1: table S14). Evaluation of these two models in the 30% blind test set revealed similar HR and C-index values of HR=3.49 (95%CI 2.73–4.47) & C=0.77 (95%CI 0.73–0.81) and HR=4.54 (95%CI 3.53–5.85) & C=0.83 (95%CI 0.80–0.86), respectively (Additional File 1: table S14). As a benchmark, chronological age alone achieved C-index values of 0.73 (0.71–0.76) and 0.76 (0.71–0.80) in training and test sets, respectively (Additional File 1: table S14). Thus, although the association with all-cause mortality is dominated by age, the naïve CD4+T-cell fraction does contribute to an improved predictive model, even when adjusted for co-morbidity (Additional File 1: table S15).

Decreased memory T-cell fractions in severe Covid-19 patients

Finally, we explored associations of immune-cell fractions with Covid-19 disease severity. Barturen et al. profiled DNAm in whole blood from 574 Covid-cases (positive at blood sample draw) and controls (negative at blood sample draw) [55]. Multivariate regression analysis adjusting for age and sex revealed multiple associations, with CD4 and CD8 memory T-cells, NK-cells, monocytes, and eosinophils all displaying reduced fractions with disease severity, while the neutrophil fraction increased (Additional File 2: fig.S17a). While the increased neutrophil fraction in severe Covid-19 cases is well-known [55, 100–102], the specific reduction in memory T-cell subsets was not previously noted by Barturen et al. We thus sought independent validation using a scRNA-Seq dataset that profiled immune cells from severe and mild Covid-19 cases as well as healthy controls (Methods) [78]. Once again we used the propeller method [77] to detect differential abundance of immune-cell subsets between severe and mild/healthy cases (Additional File 1: table S16). This confirmed the decrease in the T-cell memory fraction in severe cases (Additional File 2: fig.S17b-d). Of note, an increased T-cell memory fraction is a hallmark of recovery from severe Covid-19 disease [102]. Thus, these data demonstrate the ability of our DNAm reference matrix to find subtle immune-cell shifts with Covid-19 severity.

Discussion

This work contributes a novel DNAm reference matrix defined over 12 immune-cell types, which is valid for both Illumina and whole-genome-bisulfite sequencing DNAm data. Using this DNAm reference matrix, we performed a large meta-analysis of cell-type fractions in blood, in order to comprehensively map associations between these 12 immune-cell fractions and common phenotypes. The meta-analysis served two purposes. First, as a means of further validating our DNAm reference matrix by testing for known associations between immune-cell fractions and a broad range of phenotypes. For instance, further validating the derived fractions for naïve T-cell subsets, we confirmed the reduced naïve to mature CD8+ and CD4+ T-cell ratio with age [6, 84–87]. The increased neutrophil fraction in severe Covid-19 patients is another well-known observation [55, 100–102], which we correctly retrieved, thus further validating the neutrophil component of our DNAm reference matrix. Our meta-analysis also established that the eosinophil count does not change with age but that it is higher in males compared with females, consistent with a previous large eosinophil count study [94]. The higher NK-cell fraction in males compared to females is validated by several reports [50, 65] and a study that used single-cell RNA sequencing to compare immune-cell fractions between males and females [66]. There is also strong prior evidence for an increased NK fraction with age [38–40], and in individuals undergoing frequent physical exercise [103, 104], further validating the NK component of our reference matrix. The increased monocyte fraction in males compared to females has also been previously observed [105]. In the case of age, sex, and Covid-19 disease severity, associations inferred from the DNAm data were strongly validated using independent scRNA-Seq data.

The meta-analysis also revealed a number of biologically and clinically significant insights and connections, which were not previously known, or for which prior evidence was scarce or controversial. For instance, there was little prior evidence for an increased naïve CD4+ and T-regulatory cell fractions in women compared to men, except for one study by Bergstedt et al. [95]. A recent review reported contradictory findings for regulatory T-cells in mice [50], and two small-sized studies reported increased T-regulatory counts [65] and naïve T-cell fractions [38] in males. In light of this controversy, our meta-analysis serves to unequivocally demonstrate that the naïve and regulatory T-cell fractions are higher in women, a result that we further validated using independent scRNA sequencing data. Of note, in females, naïve CD4+T-cells preferentially produce IFNγ upon stimulation, whereas in males they produce more IL-17 [106]. Correspondingly, IFNγ production has been found to be higher in females [107]. Given IFNγ's major role in activating anticancer immunity, it is therefore not implausible that this subtle increase in the naïve CD4+T-cell fraction could contribute to the well-known overall lower cancer-incidence in women [108].

Another insight, which was seen in 3 cohorts with available BMI information (TZH, FCE, and TD7k), is the increased memory CD4+T-cell fraction in obese individuals, consistent with a previous study [109] and another reporting an increased circulating CD4+T-cell frequency with increased BMI [110]. The increased

Luo *et al. Genome Medicine*   (2023) 15:59

Page 19 of 24

CD4 + T-cell memory fraction in obese individuals could reflect increased activation due to adipocyte antigen-presenting cells [110] and inflammation [111] within adipose tissue. Another interesting connection was centered on the NK fraction, a key component of anticancer immunity, which displayed decreases with cancer risk factors such as smoking, obesity, and stress levels, while it increased with cancer-preventive factors such as exercise and hours of sleep, all these associations being derived from multivariate models that included age and sex. On the other hand, the NK fraction increased with age and was higher in males compared to females. These data clearly indicate the importance of recording all epidemiological factors in individual studies, as multiple factors can impinge on measured cell-type fractions.

Another noteworthy insight was the observation of an increased eosinophil count with age in the blood of childhood cancer survivors, when this increase was not evident in any of the other 20 cohorts. This observation could be significant as childhood cancer survivors are known to be at a much higher risk of developing heart disease [53, 112], and one particular rare condition that can lead to a range of cardiovascular disease manifestations is eosinophilic myocarditis [113], a condition associated with elevated eosinophil counts. Thus, although the elevated risk of heart disease is more likely related to long-term effects of cancer treatment, it is nevertheless also plausible that the increased cardiovascular disease risk could be driven in part by an age-associated increase in eosinophils. On the other hand, we also express caution because the childhood cancer study profiled PBMCs, which is depleted for granulocytes, and so the observed association between eosinophil counts and age could also be due to differences in mononuclear cell isolation efficiency.

Of note, the meta-analyses were performed using directional Stouffer, and separately also with a fixed effects inverse variance model, both methods yielding highly congruent and significant *P*-values. Heterogeneity, as measured by the I^2 measure, was high for effectively all immune-cell types, highlighting the diverse nature of the underlying cohorts. Although we did adjust for study and cohort-specific biases wherever possible, unknown factors that influence cell counts such as cytomegalovirus infection [95] may well contribute to such heterogeneity, posing obvious limitations to our analysis. Although a high level of heterogeneity would appear to justify a random effects (RE) model, our data indicates significant deviations from the Gaussian assumption underlying the RE model. Indeed, a RE model almost failed to predict the decrease of naïve CD8T-cell fraction with age, when this immunosenescence signature is a well-known biological fact, which we observed in all of the 21 cohorts that entered the meta-analysis for age. Overall, our data highlights the importance of performing meta-analyses, not only to increase power, but to identify recurrent associations despite the underlying heterogeneity and unknown confounders.

By merging cell-type fractions from different cohorts together, we were also able to establish that the naïve CD8 + T-cell fraction decreases non-linearly with age, displaying a more pronounced decrease in the age range 20 to 40 compared to mid-life (40–60 years), with the decrease being much less noticeable beyond 65 years of age. In contrast, the naïve CD4 + T-cell fraction displayed a linear decrease throughout life. Elucidating the biological basis of these differences could have important ramifications for our understanding of the aging immune system. Interestingly, of all immune-cell fractions, the naïve and memory CD4 + T-cell fractions displayed the strongest association with all-cause mortality, with increased fractions associated with a significantly reduced risk. These associations were not only independent of all major disease risk factors (age, sex, ethnicity, smoking, BMI, and alcohol consumption) but also independent of baseline co-morbidity. Correspondingly, the naïve CD4 + T-cell fraction also displayed significant associations with the risk of developing specific conditions, including T2D, COPD, and CVD. This is consistent with reports of a reduced naïve CD4 + T-cell count in T2D patients [114]. Of note, the association of a reduced naïve C4T-cell count with increased risk for all-cause mortality was also seen in the Lothian Birth Cohorts [115] and has been explicitly demonstrated with measured cell counts in specific clinical subgroups (e.g., hemodialysis patients [116]). The NK fraction was associated with a reduced risk of all-cause mortality as well as COPD and T2D. The reduced NK fraction with T2D risk is consistent with a study comparing cell counts of T2D patients to controls [117]. An increased neutrophil fraction was associated with all-cause mortality and risk for T2D and COPD. This is consistent with a recent cell count study in type 2 diabetes [118] and a study demonstrating an increased neutrophil-to-lymphocyte ratio (NLR) in COPD [119]. An increased NLR has also been recently associated with an increased all-cause mortality [120], with HRs similar to those reported here. Considering individual health outcomes was important: for instance, while the memory B-cell fraction did not correlate with all-cause mortality, an increase in this fraction was associated with an increased risk of cancer and simultaneously with a reduced risk of T2D. Although these specific associations do not seem to have been previously reported, it is noteworthy that an increased memory B-cell fraction could be associated with reduced anti-tumor immunity through increased release of pro-tumorigenic factors such as IL-10 or TGF-beta [121].

Luo *et al. Genome Medicine*   (2023) 15:59

Page 20 of 24

Conclusions

In summary, this study has comprehensively mapped associations between immune-cell fractions and common phenotypes at an unprecedented high resolution of 12 immune-cell subtypes in blood, revealing many important associations with factors such as age, sex, smoking, obesity, and health outcomes. The DNAm reference matrix encompassing 12 immune-cell subtypes that we present here has been extensively validated, including WGBS data, and is made freely available as part of our EpiDISH Bioconductor R-package. We envisage that similar high-resolution meta-analyses performed in tissues other than blood using tools such as EpiSCORE [22, 24] could help discern changes in cell-type composition that are important predictors or contributors of disease and disease risk.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13073-023-01211-5.

Additional file 1. An excel spreadsheet containing all Supplementary Tables.

Additional file 2. A pdf Supplementary Information file. Contains all Supplementary.

Availability of data and materials
The following DNAm datasets analyzed here are publicly available from the NCBI GEO website under the accession numbers GSE40279 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40279) [67, 122], GSE42861 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861) [5, 123], GSE50660 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50660) [68, 124], GSE106648 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106648) [52, 125], GSE169156 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169156) [53, 126], GSE210256 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210256, a SuperSeries object containing GSE210254 and GSE210255) [54, 127], GSE179325 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179325) [55, 128], GSE147740 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147740) [56, 129], GSE117860 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117860) [57, 130], GSE87648 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87648) [58, 131], GSE84727 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84727) [59, 132], GSE80417 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80417) [60, 133], GSE72680 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72680) [61, 134], GSE61151 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61151) [62, 135], GSE87571 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87571) [63, 136], GSE55763 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55763) [64, 137].

The MRC1946 DNAm data [18] is only available by submitting data requests to mrclha.swiftinfo@ucl.ac.uk; see full policy at http://www.nshd.mrc.ac.uk/data.aspx. Managed access is in place for this study to ensure that use of the data are within the bounds of consent given previously by participants, and to safeguard any potential threat to anonymity since the participants are all born in the same week. The FCE DNAm data [69] is available from the European Genome Archive (EGA) under accession number EGAS00001005626 (https://ega-archive.org/search-results.php?query=EGAS00001005626) [138]. The Illumina EPIC DNAm data for the TZH cohort [66] can be viewed at NODE under accession number OEP000260 (https://www.biosino.org/node/project/detail/OEP000260) [139], and accessed by submitting a request for data access. Data usage shall be in full compliance with the Regulations on Management of Human Genetic Resources in China. The TD7k, MGB, and UCLA datasets are available upon request to TruDiagnostic Inc. (varun@trudiagnostic.com). In order to protect data privacy of the individuals represented in this cohort, individual applications will be reviewed by TD and in case TD is willing to share data, a data sharing agreement will be set up. The scRNA-Seq data from Yazar et al. [76] is publicly available from the NCBI GEO website under the accession number GSE196830 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196830) [140]. The scRNA-Seq data from Liao et al. [78] is publicly available from the NCBI GEO website under the accession number GSE145926 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145926) [141]. The two 12 blood cell-type DNAm reference matrices are available in Additional File 1: SI tables S1 and S2, and have also been incorporated into the EpiDISH BioC R-package [142] which includes tools for cell-type fraction estimation.

Declarations

Author details
[1]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China. [2]TruDiagnostics, 881 Corporate Dr., Lexington, KY 40503, USA. [3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA. [4]Institute for Health and Sport (iHeS), Victoria University, Footscray, VIC 3011, Australia. [5]PhysioAge LLC, 30 Central Park South / Suite 8A, New York, NY 10019, USA. [6]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, USA. [7]Australian Regenerative Medicine Institute, Monash University, Clayton, VIC 3800, Australia.

References
1. Lappalainen T, Greally JM. Associating cellular epigenetic models with human phenotypes. Nat Rev Genet. 2017;18:441–51.
2. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. Nat Rev Genet. 2018;19:129–47.
3. Ali HR, Provenzano E, Dawson SJ, Blows FM, Liu B, Shah M, Earl HM, Poole CJ, Hiller L, Dunn JA, et al. Association between CD8+ T-cell infiltration and breast cancer survival in 12,439 patients. Ann Oncol. 2014;25:1536–43.

Luo *et al. Genome Medicine*   (2023) 15:59

Page 21 of 24

4.   Crooke SN, Ovsyannikova IG, Poland GA, Kennedy RB. Immunosenescence: a systems-level overview of immune cell biology and strategies for improving vaccine responses. Exp Gerontol. 2019;124:110632.

5.   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.

6.   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.

7.   Koestler DC, Marsit CJ, Christensen BC, Accomando W, Langevin SM, Houseman EA, Nelson HH, Karagas MR, Wiencke JK, Kelsey KT. Peripheral blood immune cell methylation profiles are associated with nonhematopoietic cancers. Cancer Epidemiol Biomarkers Prev. 2012;21:1293–302.

8.   Langevin SM, Houseman EA, Accomando WP, Koestler DC, Christensen BC, Nelson HH, Karagas MR, Marsit CJ, Wiencke JK, Kelsey KT. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. Epigenetics. 2014;9:884–95.

9.   Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011;12:529–41.

10. Fasanelli F, Baglietto L, Ponzi E, Guida F, Campanella G, Johansson M, Grankvist K, Johansson M, Assumma MB, Naccarati A, et al. Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. Nat Commun. 2015;6:10192.

11. Talukdar FR, Escobar Marcillo DI, Laskar RS, Novoloaca A, Cuenin C, Sbraccia P, Nistico L, Guglielmi V, Gheit T, Tommasino M, et al. Bariatric surgery-induced weight loss and associated genome-wide DNA-methylation alterations in obese individuals. Clin Epigenetics. 2022;14:176.

12. van Eijk KR, de Jong S, Strengman E, Buizer-Voskamp JE, Kahn RS, Boks MP, Horvath S, Ophoff RA. Identification of schizophrenia-associated loci by combining DNA methylation and gene expression data from whole blood. Eur J Hum Genet. 2015;23:1106–10.

13. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R, et al. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5:e1000602.

14. Dekkers KF, van Iterson M, Slieker RC, Moed MH, Bonder MJ, van Galen M, Mei H, Zhernakova DV, van den Berg LH, Deelen J, et al. Blood lipids influence DNA methylation in circulating cells. Genome Biol. 2016;17:138.

15. Wahl S, Drong A, Lehne B, Loh M, Scott WR, Kunze S, Tsai PC, Ried JS, Zhang W, Yang Y, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. Nature. 2017;541:81–6.

16. Maity AK, Stone TC, Ward V, Webster AP, Yang Z, Hogan A, McBain H, Duku M, Ho KMA, Wolfson P, et al. Novel epigenetic network biomarkers for early detection of esophageal cancer. Clin Epigenetics. 2022;14:23.

17. Wong YT, Tayeb MA, Stone TC, Lovat LB, Teschendorff AE, Iwasiow R, Craig JM. A comparison of epithelial cell content of oral samples estimated using cytology and DNA methylation. Epigenetics. 2022;17:327–34.

18. Teschendorff AE, Yang Z, Wong A, Pipinikas CP, Jiao Y, Jones A, Anjum S, Hardy R, Salvesen HB, Thirlwell C, et al. Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer. JAMA Oncol. 2015;1:476–85.

19. Beck S. Taking the measure of the methylome. Nat Biotechnol. 2010;28:1026–8.

20. 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 Bioinform. 2012;13:86.

21. 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 Bioinform. 2017;18:105.

22. Zhu T, Liu J, Beck S, Pan S, Capper D, Lechner M, Thirlwell C, Breeze CE, Teschendorff AE. A pan-tissue DNA methylation atlas enables in silico decomposition of human tissue methylomes at cell-type resolution. Nat Methods. 2022;19:296–306.

23. 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

24. Teschendorff AE, Zhu T, Breeze CE, Beck S. EPISCORE: cell type deconvolution of bulk tissue DNA methylomes from single-cell RNA-Seq data. Genome Biol. 2020;21:221.

25. 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 HumanMethylationEPIC BeadArray. Genome Biol. 2018;19:64.

26. Koestler DC, Jones MJ, Usset J, Christensen BC, Butler RA, Kobor MS, Wiencke JK, Kelsey KT. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). BMC Bioinformatics. 2016;17:120.

27. 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.

28. 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.

29. Loyfer N, Magenheim 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.

30. Stunnenberg HG. International Human Epigenome C, Hirst M: The international human epigenome consortium: a blueprint for scientific collaboration and discovery. Cell. 2016;167:1145–9.

31. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014;30:1363–9.

32. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29:189–96.

33. Teschendorff AE, Zhuang J, Widschwendter M. Independent surrogate variable analysis to deconvolve confounding factors in large-scale microarray profiling studies. Bioinformatics. 2011;27:1496–505.

34. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:Article3.

35. 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.

36. Reynolds LM, Taylor JR, Ding J, Lohman K, Johnson C, Siscovick D, Burke G, Post W, Shea S, Jacobs DR Jr, et al. Age-related variations in the methylome associated with gene expression in human monocytes and T cells. Nat Commun. 2014;5:5366.

37. Chen L, Ge B, Casale FP, Vasquez L, Kwan T, Garrido-Martin D, Watt S, Yan Y, Kundu K, Ecker S, et al. Genetic drivers of epigenetic and transcriptional variation in human immune cells. Cell. 2016;167(1398–1414):e1324.

38. Coit P, Yalavarthi S, Ognenovski M, Zhao W, Hasni S, Wren JD, Kaplan MJ, Sawalha AH. Epigenome profiling reveals significant DNA demethylation of interferon signature genes in lupus neutrophils. J Autoimmun. 2015;58:59–66.

39. Nestor CE, Barrenas F, Wang H, Lentini A, Zhang H, Bruhn S, Jornsten R, Langston MA, Rogers G, Gustafsson M, Benson M. DNA methylation changes separate allergic patients from healthy controls and may reflect altered CD4+ T-cell population structure. PLoS Genet. 2014;10:e1004059.

40. Shade DC, Park HJ, Hausman DB, Hohos N, Meagher RB, Kauwell GPA, Kilaru V, 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. Int J Vitam Nutr Res. 2017;87:271–8.

41. Limbach M, Saare M, Tserel L, Kisand K, Eglit T, Sauer S, Axelsson T, Syvanen AC, Metspalu A, Milani L, Peterson P. Epigenetic profiling in CD4+ and CD8+ T cells from Graves' disease patients reveals changes in genes associated with T cell receptor signaling. J Autoimmun. 2016;67:46–56.

Luo *et al. Genome Medicine*   (2023) 15:59

Page 22 of 24

42. Mamrut S, Avidan N, Staun-Ram E, Ginzburg E, Truffault F, Berrih-Aknin S, Miller A. Integrative analysis of methylome and transcriptome in human blood identifies extensive sex- and immune cell-specific differentially methylated regions. Epigenetics. 2015;10:943–57.

43. Absher DM, Li X, Waite LL, Gibson A, Roberts K, Edberg J, Chatham WW, Kimberly RP. Genome-wide DNA methylation analysis of systemic lupus erythematosus reveals persistent hypomethylation of interferon genes and compositional changes to CD4+ T-cell populations. PLoS Genet. 2013;9:e1003678.

44. Tserel L, Kolde R, Limbach M, Tretyakov K, Kasela S, Kisand K, Saare M, Vilo J, Metspalu A, Milani L, Peterson P. Age-related profiling of DNA methylation in CD8+ T cells reveals changes in immune response and transcriptional regulator genes. Sci Rep. 2015;5:13107.

45. 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.

46. Reinius 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.

47. 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.

48. van Rooij J, Mandaviya PR, Claringbould A, Felix JF, van Dongen J, Jansen R, Franke L. consortium B, t Hoen PAC, Heijmans B, van Meurs JBJ: Evaluation of commonly used analysis strategies for epigenome- and transcriptome-wide association studies through replication of large-scale population studies. Genome Biol. 2019;20:235.

49. Wu MC, Joubert BR, Kuan PF, Haberg SE, Nystad W, Peddada SD, London SJ. A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics. 2014;9:318–29.

50. Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerstrom-Billai F, Jagodic M, Sundberg CJ, Ekstrom TJ, Teschendorff AE, Tegner J, Gomez-Cabrero D. An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform. Epigenetics. 2013;8:333–46.

51. Welsh H, Batalha C, Li W, Mpye KL, Souza-Pinto NC, Naslavsky MS, Parra EJ. A systematic evaluation of normalization methods and probe replicability using infinium EPIC methylation data. Clin Epigenetics. 2023;15:41.

52. Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, James T, Ewing E, Lindén M, Górnikiewicz B, et al. DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. Nat Commun. 2018;9:2397.

53. Song N, Hsu CW, Pan H, Zheng Y, Hou L, Sim JA, Li Z, Mulder H, Easton J, Walker E, et al. Persistent variations of blood DNA methylation associated with treatment exposures and risk for cardiometabolic outcomes in long-term survivors of childhood cancer in the St. Jude Lifetime Cohort Genome Med. 2021;13:53.

54. Shang L, Zhao W, Wang YZ, Li Z, Choi JJ, Kho M, Mosley TH, Kardia SLR, Smith JA, Zhou X. meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. Nat Commun. 2023;14:2711.

55. Barturen G, Carnero-Montoro E, Martinez-Bueno M, Rojo-Rello S, Sobrino B, Porras-Perales O, Alcantara-Dominguez C, Bernardo D, Alarcon-Riquelme ME. Whole blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. Nat Commun. 2022;13:4597.

56. Robinson O, Chadeau Hyam M, Karaman I, Climaco Pinto R, Ala-Korpela M, Handakas E, Fiorito G, Gao H, Heard A, Jarvelin MR, et al. Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. Aging Cell. 2020;19:e13149.

57. Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, Hulgan T, Bryant KJ, Zhao H, Krystal JH, et al. Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. Clin Epigenetics. 2018;10:155.

58. Ventham NT, Kennedy NA, Adams AT, Kalla R, Heath S, O'Leary KR, Drummond H, consortium IB, consortium IC, Wilson DC, et al: Integrative epigenome-wide analysis demonstrates that DNA methylation

59. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, St Clair D, Mustard C, Breen G, Therman S, et al. An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. Genome Biol. 2016;17:176.

60. Hannon E, Dempster EL, Mansell G, Burrage J, Bass N, Bohlken MM, Corvin A, Curtis CJ, Dempster D, Di Forti M, et al. DNA methylation meta-analysis reveals cellular alterations in psychosis and markers of treatment-resistant schizophrenia. Elife. 2021;10:e58430.

61. 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.

62. Flanagan JM, Brook MN, Orr N, Tomczyk K, Coulson P, Fletcher O, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow A, et al. Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. Cancer Epidemiol Biomarkers Prev. 2015;24:221–9.

63. Johansson A, Enroth S, Gyllensten U. Continuous aging of the human DNA methylome throughout the human lifespan. PLoS ONE. 2013;8:e67378.

64. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan ST, Afzal U, Scott J, Jarvelin MR, Elliott P, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol. 2015;16:37.

65. Voisin S, Jacques M, Landen S, Harvey NR, Haupt LM, Griffiths LR, Gancheva S, Ouni M, Jahnert M, Ashton KJ, et al. Meta-analysis of genome-wide DNA methylation and integrative omics of age in human skeletal muscle. J Cachexia Sarcopenia Muscle. 2021;12:1064–78.

66. 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.

67. 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. Mol Cell. 2013;49:359–67.

68. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Vinuela A, Grundberg E, Nelson CP, Meduri E, et al. Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. Epigenetics. 2014;9:1382–96.

69. Barrett JE, Herzog C, Kim YN, Bartlett TE, Jones A, Evans I, Cibula D, Zikan M, Bjorge L, Harbeck N, et al. Susceptibility to hormone-mediated cancer is reflected by different tick rates of the epithelial and general epigenetic clock. Genome Biol. 2022;23:52.

70. Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. Aging (Albany NY). 2018;10:3541–57.

71. Singaravelu G, Harley CB, Raffaele JM, Sudhakaran P, Suram A. Double-blind, placebo-controlled, randomized clinical trial demonstrates telomerase activator TA-65 decreases immunosenescent CD8 + CD28 - T cells in humans. OBM Geriatrics. 2021;5:1–26.

72. Zheng SC, Breeze CE, Beck S, Dong D, Zhu T, Ma L, Ye W, Zhang G, Teschendorff AE. EpiDISH web server: epigenetic dissection of intra-sample-heterogeneity with online GUI. Bioinformatics. 2019;36(6):1950–1.

73. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. Res Synthesis Meth. 2010;1:97–111.

74. Balduzzi S, Rucker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health. 2019;22:153–60.

75. Zeileis A, Hothorn T. Diagnostic checking in regression relationships. R News. 2002;2:7–10.

76. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, Lu Q, Rowson A, Taylor TRP, Clarke L, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science. 2022;376:eabf3041.

may mediate genetic risk in inflammatory bowel disease. Nat Commun 2016, 7:13507.

Luo *et al. Genome Medicine*   (2023) 15:59

Page 23 of 24

77. Phipson B, Sim CB, Porrello ER, Hewitt AW, Powell J, Oshlack A. propeller: testing for differences in cell type proportions in single cell data. Bioinformatics. 2022;38:4720–6.

78. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26:842–4.

79. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36:411–20.

80. Castro VM, Gainer V, Wattanasin N, Benoit B, Cagan A, Ghosh B, Goryachev S, Metta R, Park H, Wang D, et al. The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. J Am Med Inform Assoc. 2022;29:643–51.

81. Nalichowski R, Keogh D, Chueh HC, Murphy SN. Calculating the benefits of a research patient data repository. AMIA Annu Symp Proc. 2006;2006:1044.

82. Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. Nat Mater. 2019;18:422–7.

83. Stunnenberg HG, International Human Epigenome C, Hirst M. the international human epigenome consortium: a blueprint for scientific collaboration and discovery. Cell. 1897;2016:167.

84. 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.

85. 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.

86. Hu B, Jadhav RR, Gustafson CE, Le Saux S, Ye Z, Li X, Tian L, Weyand CM, Goronzy JJ. Distinct age-related epigenetic signatures in CD4 and CD8 T Cells. Front Immunol. 2020;11:585168.

87. Li M, Yao D, Zeng X, Kasakovski D, Zhang Y, Chen S, Zha X, Li Y, Xu L. Age related human T cell subset evolution and senescence. Immun Ageing. 2019;16:24.

88. Elyahu Y, Hekselman I, Eizenberg-Magar I, Berner O, Strominger I, Schiller M, Mittal K, Nemirovsky A, Eremenko E, Vital A, et al. Aging promotes reorganization of the CD4 T cell landscape toward extreme regulatory and effector phenotypes. Sci Adv. 2019;5:eaaw8330.

89. Maity AK, Hu X, Zhu T, Teschendorff AE. Inference of age-associated transcription factor regulatory activity changes in single cells. Nat Aging. 2022;2:548–61.

90. Camous X, Pera A, Solana R, Larbi A. NK cells in healthy aging and age-associated diseases. J Biomed Biotechnol. 2012;2012:195956.

91. Gayoso I, Sanchez-Correa B, Campos C, Alonso C, Pera A, Casado JG, Morgado S, Tarazona R, Solana R. Immunosenescence of human natural killer cells. J Innate Immun. 2011;3:337–43.

92. Borrego F, Alonso MC, Galiani MD, Carracedo J, Ramirez R, Ostos B, Pena J, Solana R. NK phenotypic markers and IL2 response in NK cells from elderly people. Exp Gerontol. 1999;34:253–65.

93. van Beek AA, Fransen F, Meijer B, de Vos P, Knol EF, Savelkoul HFJ. Aged mice display altered numbers and phenotype of basophils, and bone marrow-derived basophil activation, with a limited role for aging-associated microbiota. Immun Ageing. 2018;15:32.

94. Hartl S, Breyer MK, Burghuber OC, Ofenheimer A, Schrott A, Urban MH, Agusti A, Studnicka M, Wouters EFM, Breyer-Kohansal R. Blood eosinophil count in the general population: typical values and potential confounders. Eur Respir J. 2020;55:1901874.

95. Bergstedt J, Azzou SAK, Tsuo K, Jaquaniello A, Urrutia A, Rotival M, Lin DTS, MacIsaac JL, Kobor MS, Albert ML, et al. The immune factors driving DNA methylation variation in human blood. Nat Commun. 2022;13:5895.

96. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016;16:626–38.

97. Nakata A, Takahashi M, Irie M, Fujioka Y, Haratani T, Araki S. Relationship between cumulative effects of smoking and memory CD4+ T lymphocyte subpopulations. Addict Behav. 2007;32:1526–31.

98. Tollerud DJ, Clark JW, Brown LM, Neuland CY, Mann DL, Pankiw-Trost LK, Blattner WA, Hoover RN. Association of cigarette smoking with decreased numbers of circulating natural killer cells. Am Rev Respir Dis. 1989;139:194–8.

99. Jung YS, Park JH, Park DI, Sohn CI, Lee JM, Kim TI. Impact of smoking on human natural killer cell activity: a large cohort study. J Cancer Prev. 2020;25:13–20.

100. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan. China Lancet. 2020;395:497–506.

101. Liu J, Li S, Liu J, Liang B, Wang X, Wang H, Li W, Tong Q, Yi J, Zhao L, et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. EBioMedicine. 2020;55:102763.

102. Odak I, Barros-Martins J, Bosnjak B, Stahl K, David S, Wiesner O, Busch M, Hoeper MM, Pink I, Welte T, et al. Reappearance of effector T cells is associated with recovery from COVID-19. EBioMedicine. 2020;57:102885.

103. Llavero F, Alejo LB, Fiuza-Luces C, Lopez Soto A, Valenzuela PL, Castillo-Garcia A, Morales JS, Fernandez D, Aldazabal IP, Ramirez M, et al. Exercise training effects on natural killer cells: a preliminary proteomics and systems biology approach. Exerc Immunol Rev. 2021;27:125–41.

104. Shephard RJ, Shek PN. Effects of exercise and training on natural killer cell counts and cytolytic activity: a meta-analysis. Sports Med. 1999;28:177–95.

105. Huang Z, Chen B, Liu X, Li H, Xie L, Gao Y, Duan R, Li Z, Zhang J, Zheng Y, Su W. Effects of sex and aging on the immune cell landscape as assessed by single-cell transcriptomic analysis. Proc Natl Acad Sci U S A. 2021;118:e2023216118.

106. Zhang MA, Rego D, Moshkova M, Kebir H, Chruscinski A, Nguyen H, Akkermann R, Stanczyk FZ, Prat A, Steinman L, Dunn SE. Peroxisome proliferator-activated receptor (PPAR)alpha and -gamma regulate IFNgamma and IL-17A production by human T cells in a sex-specific way. Proc Natl Acad Sci U S A. 2012;109:9505–10.

107. Roberts CW, Walker W, Alexander J. Sex-associated hormones and immunity to protozoan parasites. Clin Microbiol Rev. 2001;14:476–88.

108. Dorak MT, Karpuzoglu E. Gender differences in cancer susceptibility: an inadequately addressed issue. Front Genet. 2012;3:268.

109. Yang H, Youm YH, Vandanmagsar B, Rood J, Kumar KG, Butler AA, Dixit VD. Obesity accelerates thymic aging. Blood. 2009;114:3803–12.

110. Zhao Y, Lin L, Li J, Xiao Z, Chen B, Wan L, Li M, Wu X, Hin Cho C, Shen J. CD4(+) T cells in obesity and obesity-associated diseases. Cell Immunol. 2018;332:1–6.

111. Zou J, Lai B, Zheng M, Chen Q, Jiang S, Song A, Huang Z, Shi P, Tu X, Wang D, et al. CD4+ T cells memorize obesity and promote weight regain. Cell Mol Immunol. 2018;15:630–9.

112. Dixon SB, Bjornard KL, Alberts NM, Armstrong GT, Brinkman TM, Chemaitilly W, Ehrhardt MJ, Fernandez-Pineda I, Force LM, Gibson TM, et al. Factors influencing risk-based care of the childhood cancer survivor in the 21st century. CA Cancer J Clin. 2018;68:133–52.

113. Al Ali AM, Straatman LP, Allard MF, Ignaszewski AP. Eosinophilic myocarditis: case series and review of literature. Can J Cardiol. 2006;22:1233–7.

114. Xia C, Rao X, Zhong J. Role of T lymphocytes in type 2 diabetes and diabetes-associated inflammation. J Diabetes Res. 2017;2017:6494795.

115. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, Gibson J, Henders AK, Redmond P, Cox SR, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome Biol. 2015;16:25.

116. Xiang F, Chen R, Cao X, Shen B, Chen X, Ding X, Zou J. Premature aging of circulating T cells predicts all-cause mortality in hemodialysis patients. BMC Nephrol. 2020;21:271.

117. Menart-Houtermans B, Rutter R, Nowotny B, Rosenbauer J, Koliaki C, Kahl S, Simon MC, Szendroedi J, Schloot NC, Roden M. German Diabetes Study G: Leukocyte profiles differ between type 1 and type 2 diabetes and are associated with metabolic phenotypes: results from the German Diabetes Study (GDS). Diabetes Care. 2014;37:2326–33.

118. Ratter-Rieck JM, Maalmi H, Trenkamp S, Zaharia OP, Rathmann W, Schloot NC, Strassburger K, Szendroedi J, Herder C, Roden M, et al. Leukocyte counts and T-cell frequencies differ between novel subgroups of diabetes and are associated with metabolic parameters and biomarkers of inflammation. Diabetes. 2021;70:2652–62.

119. Gunay E, Sarinc Ulasli S, Akar O, Ahsen A, Gunay S, Koyuncu T, Unlu M. Neutrophil-to-lymphocyte ratio in chronic obstructive pulmonary disease: a retrospective study. Inflammation. 2014;37:374–80.

Luo *et al. Genome Medicine*   (2023) 15:59

Page 24 of 24

120. Chen Y, Wang W, Zeng L, Mi K, Li N, Shi J, Yang S. Association between neutrophil-lymphocyte ratio and all-cause mortality and cause-specific mortality in US adults, 1999–2014. Int J Gen Med. 2021;14:10203–11.

121. Downs-Canner SM, Meier J, Vincent BG, Serody JS. B cell function in the tumor microenvironment. Annu Rev Immunol. 2022;40:169–93.

122. 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. GSE40279, NCBI Gene Expression Omnibus GEO; 2013. https://www.ncbi.nlm.nih.gov/query/acc.cgi?acc=GSE40279.

123. 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. GSE42861, NCBI Gene Expression Omnibus GEO; 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42861.

124. Tsaprouni LG, Yang TP, Bell J, Dick KJ, Kanoni S, Nisbet J, Viñuela A, Grundberg E, Nelson CP, Meduri E, et al: Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation. GSE50660, NCBI Gene Expression Omnibus GEO; 2014. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE50660.

125. Kular L, Liu Y, Ruhrmann S, Zheleznyakova G, Marabita F, Gomez-Cabrero D, James T, Ewing E, Lindén M, Górnikiewicz B, et al: DNA methylation as a mediator of HLA-DRB1*15:01 and a protective variant in multiple sclerosis. GSE106648, NCBI Gene Expression Omnibus GEO; 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE106648.

126. Song N, Hsu CW, Pan H, Zheng Y, Hou L, Sim JA, Li Z, Mulder H, Easton J, Walker E, et al: Persistent variations of blood DNA methylation associated with treatment exposures and risk for cardiometabolic outcomes in long-term survivors of childhood cancer in the St. Jude Lifetime Cohort. In Genome Med. GSE169156, NCBI Gene Expression Omnibus GEO; 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE169156.

127. Shang L, Zhao W, Wang YZ, Li Z, Choi JJ, Kho M, Mosley TH, Kardia SLR, Smith JA, Zhou X: meQTL mapping in the GENOA study reveals genetic determinants of DNA methylation in African Americans. GSE210256, NCBI Gene Expression Omnibus GEO; 2023. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210256.

128. Barturen G, Carnero-Montoro E, Martínez-Bueno M, Rojo-Rello S, Sobrino B, Porras-Perales Ó, Alcántara-Domínguez C, Bernardo D, Alarcón-Riquelme ME: Whole blood DNA methylation analysis reveals respiratory environmental traits involved in COVID-19 severity following SARS-CoV-2 infection. GSE179325, NCBI Gene Expression Omnibus GEO; 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179325.

129. Robinson O, Chadeau Hyam M, Karaman I, Climaco Pinto R, Ala-Korpela M, Handakas E, Fiorito G, Gao H, Heard A, Jarvelin MR, et al: Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. GSE147740, NCBI Gene Expression Omnibus GEO; 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147740.

130. Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, Hulgan T, Bryant KJ, Zhao H, Krystal JH, et al: Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. GSE117860, NCBI Gene Expression Omnibus GEO; 2018. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117860.

131. Ventham NT, Kennedy NA, Adams AT, Kalla R, Heath S, O'Leary KR, Drummond H, consortium IB, consortium IC, Wilson DC, et al: Integrative epigenome-wide analysis demonstrates that DNA methylation may mediate genetic risk in inflammatory bowel disease. GSE87648, NCBI Gene Expression Omnibus GEO; 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87648.

132. Hannon E, Dempster E, Viana J, Burrage J, Smith AR, Macdonald R, St Clair D, Mustard C, Breen G, Therman S, et al: An integrated genetic-epigenetic analysis of schizophrenia: evidence for co-localization of genetic associations and differential DNA methylation. GSE84727, NCBI Gene Expression Omnibus GEO; 2016. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE84727.

133. Hannon E, Dempster EL, Mansell G, Burrage J, Bass N, Bohlken MM, Corvin A, Curtis CJ, Dempster D, Di Forti M, et al: DNA methylation meta-analysis reveals cellular alterations in psychosis and markers

of treatment-resistant schizophrenia. vol. 10, 2021/03/02 edition. GSE80417, NCBI Gene Expression Omnibus GEO; 2021. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE80417.

134. 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. GSE72680, NCBI Gene Expression Omnibus GEO; 2019. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE72680.

135. Flanagan JM, Brook MN, Orr N, Tomczyk K, Coulson P, Fletcher O, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow A, et al: Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study. GSE61151, NCBI Gene Expression Omnibus GEO; 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE61151.

136. Johansson A, Enroth S, Gyllensten U: Continuous aging of the human DNA methylome throughout the human lifespan. GSE87571, NCBI Gene Expression Omnibus GEO; 2013. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87571.

137. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan ST, Afzal U, Scott J, Jarvelin MR, Elliott P, et al: A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. GSE55763, NCBI Gene Expression Omnibus GEO; 2015. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE55763.

138. Barrett JE, Herzog C, Kim YN, Bartlett TE, Jones A, Evans I, Cibula D, Zikan M, Bjorge L, Harbeck N, et al: Susceptibility to hormone-mediated cancer is reflected by different tick rates of the epithelial and general epigenetic clock. European Genome Archive EGAS00001005626. 2022. https://ega-archive.org/search-results.php?query=EGAS00001005626.

139. 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. 2020/09/24 edition. NODE OEP000260. 2020. https://www.biosino.org/node/project/detail/OEP000260.

140. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, Lu Q, Rowson A, Taylor TRP, Clarke L, et al: Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. GSE196830, NCBI Gene Expression Omnibus GEO; 2022. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE196830.

141. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, Cheng L, Li J, Wang X, Wang F, et al: Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. GSE145926, NCBI Gene Expression Omnibus GEO; 2020. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE145926.

142. Zheng SC, Teschendorff AE: EpiDISH: Epigenetic deconvolution of intra-sample heterogeneity. Bioconductor; 2023. https://bioconductor.org/packages/release/bioc/html/EpiDISH.html.

Publisher's Note