Supplementary Materials: Multi-modal profiling of peripheral blood cells across the human lifespan reveals distinct immune cell signatures of aging and longevity

Supplementary Methods

Experimental Procedure:

Recruitment of human subjects. Centenarian study participants were recruited via recruitment mailings based upon statewide voter registration lists throughout the United States. A healthy volunteer effect made it more likely that the enrolled centenarian sample was healthier than centenarians generally. Participants with capacity provided informed written consent and otherwise next of kin acted as a legally authorized representative in providing informed written consent on behalf of the participant. The Boston University Medical Campus Institutional Review Board (BUMC IRB) reviewed and approved this minimum risk study of centenarians and their family members. For participants asked to provide a blood sample for the creation of induced pluripotent stem cell lines, they provided informed written consent for a separate study, again reviewed and approved by the BUMC IRB that is conducted by the Center for Regenerative Medicine also based on the Boston University Medical Campus.

Processing of blood samples. For each centenarian and younger individual involved in this study, 8 mLs of peripheral blood was drawn into each of two BD Vacutainer Cell Preparation Tubes with sodium citrate (BD Biosciences catalog #362760). The tubes were centrifuged at 1,800 x g for 30 minutes at room temperature (RT) and cellular fraction was processed for cell isolation. The cell layer containing peripheral blood mononuclear cells (PBMCs) isolated by Ficoll gradient centrifugation was transferred into a sterile 15 mL conical centrifuge tube. The PBMC sample was brought to 10 mLs with sterile Dulbecco's phosphate buffered saline (DPBS, Invitrogen catalog #14190-144) and centrifuged at 300 x g for 15 min at RT. The supernatant was aspirated and the pellet resuspended in 10 mL sterile DPBS and a cell count performed via hemocytometer. The sample was centrifuged at 300 x g for 10 min at RT and the supernatant aspirated. The pellet was resuspended in chilled (4℃) resuspension medium (40% fetal bovine serum (FBS) hyclone defined, Cytiva catalog #SH30070.03 in Iscove's Modified Dulbecco's Medium (IMDM), catalog #12440053) to achieve a cell concentration of 4 x 10^6 cells/ mL. An equal amount of chilled 2X freezing medium (30% Dimethyl Sulfoxide (DMSO), Sigma catalog #D2650 in IMDM/ 40% FBS) was added to achieve a cell concentration of 2 x 10^6 cells / mL. This mixture was then aliquoted into 1.2 mL cryovials (Corning catalog #430487) at 1 mL / vial. These vials were then brought to -80℃ before being transferred to a -150℃ deep freezer.

CITE-seq of PBMCs of centenarians. PBMC samples (2 x 10^6 cells/sample) from the centenarian cohort were thawed rapidly and mixed with 15 mL StemSpan SFEM II medium (CAT#09655) with L-glutamine (1:1000 conc). These samples were then brought to 50 mL with sort buffer (2% Bovine Serum Albumin (BSA), Millipore Sigma catalog #EM-2930 in DPBS) and centrifuged for 5 min at 400 x g at RT. The supernatant was aspirated, and the pellet of PBMCs resuspended in 30 mL sort buffer and centrifuged for 5 min at 400 x g at RT. The pellet was resuspended in 2 mL StemSpan medium (+L-glutamine) and filtered through a 40 uM filter. The samples were then incubated in this medium for 1 hour at 37℃/5% CO2. Following this incubation, the samples were centrifuged for 5 min at 400 rcf at RT and the supernatant aspirated. Each pellet was then resuspended in 50 uL labeling buffer (1% BSA in PBS) and 5 uL Human TruStain FcX(Biolegend) was added. The samples were incubated at 4℃ for 10 minutes. During this incubation, the TotalSeq-C antibody pool (Biolegend) containing 1 ug of each antibody was prepared and centrifuged for 10 min at 14,000 x g at 4℃. The supernatant was then transferred and used as the antibody mix for each sample. The following TotalSeq-C antibodies were used: CD274, CD3, CD8, CD19, CD33, CD4, CD14, CD16, CD56, and CD279 (BioLegend). 20 uL of the antibody mix was added to each sample and the samples were brought to 100 uL with labeling buffer. The samples were incubated for 30 minutes at 4℃. Following this incubation, the samples were washed with 1.3 mL labeling buffer and centrifuged at 400 x g for 5 minutes at RT. This washing step was repeated for a total of 3 washes. The pellets were then resuspended in 500 uL labeling buffer with calcein blue AM (1:1000) (Thermo Fisher, catalog #C1429). 1 x 10^5 calcein blue positive cells (live cells) were then sorted on a Beckman Coulter MoFlo Astrios cell sorter. The sorted samples were centrifuged for 5 min at 400 rcf at room temperature and resuspended in 120 uL resuspension buffer (0.04% BSA in DPBS). The samples were counted via hemocytometer and diluted to 600 cells/uL with resuspension buffer.

Flow cytometry analysis. Frozen PBMCs from centenarian and younger samples were banked and thawed and following the same protocol as described above and below for the scRNA-seq. Cytometry panel design and validation, sample staining and sample acquisition were performed closely following OMIP-069 protocol[1]. Briefly, cells were stained with a live-dead dye (Live Dead Blue, Thermo Fisher), blocked with FcBlock reagent (Biolegend) and Monocyte Blocker (Biolegend), stained with the fluorescent antibody cocktail and Brilliant Buffer Plus (BD Biosciences), washed and analyzed on the Cytek Aurora spectral cytometer (Cytek Biosciences). At least 500,000 cells were recorded for each PBMC sample. Younger PBMCs and Ultracomp eBeads plus (Thermo Fisher) were used as single stain unmixing controls. Data were processed in SpectoFlo 2.2 (Cytek Biosciences) and OMIQ data analysis platform. We reproduced the exact gating strategy as shown in Alpert et al [2] and extracted cell counts and population proportions for downstream statistical analysis. The following antibodies were used to identify cell populations: CD45RA BUV395, CD16 BUV496, CD56 BUV737, CD8 BUV805, CCR7 BV421 (BD Biosciences), CD123 Super Bright 436 (Thermo Fisher), CD33 BV510, CD14 BV570, CD3 Spark Blue 550, CD19 Spark NIR 685 (Biolegend), CD4 cFluor YG584 (Cytek Biosciences).

Single cell analysis:

New England Centenarian dataset:

CITE-seq and CellRanger Preprocessing. Cellular Indexing and epitopes sequencing (CITE-seq) was performed on the 7 centenarians and 2 younger age individuals using a commercial droplet-based platform (10x Chromium). We constructed 5' gene expression libraries (GEX), as well as surface protein libraries (antibody derived tags, ADT) following the manufacturer's user guide. These libraries were sequenced on two runs of an Illumina NextSeq 2000 instrument generating 438 and 535 million reads respectively. Raw sequencing files were converted to fastq and demultiplexed using bcl2fastq v.2.20 and Cellranger v.3.0.2. Counts for the expression and antibody capture libraries were derived by simultaneously mapping the respective fastq files to the human genome (GRCh38) and to the feature reference of the TotalSeq-C antibodies used using the corresponding parameters in cellranger count (v.3.0.2). This pipeline includes the alignment, barcode and UMI counting

Filtering, PCA Analysis, Batch Correction, and Clustering. After processing the samples through CellRanger, we performed filtering, normalization, and principal component analysis using Seurat v.3 [3].

First, we performed quality control steps based on the number of genes and UMIs detected per cell, and percent of mitochondrial genes expressed per cell. To remove poor quality cells with low RNA content, we removed cells with less than 200 genes detected. To filter out outlier cells and doublets, we filtered out cells with greater than 3,000 detected genes, as well as cells with greater than 15,000 UMIs. To account for cells that are damaged or dying, we removed cells with greater than 15 percent mitochondrial counts expressed.

After filtering, we normalized the RNA-level expression data for each cell to compare gene expression between sample cells; Gene counts for each cell were normalized by total expression, multiplied by a scale factor of 10,000 and transformed to a log scale. We normalized the protein-level expression data by applying a centered log ratio (CLR) transformation for each cell to account for differences in total protein ADT counts that make up each cell.

Further downstream analyses were performed on the RNA-level expression data. PCA based on the top 2000 highly variable genes was performed for dimensionality reduction, and the top 20 significant PCs were selected that explain the most variability in the data. The top significant PCs were used as an input for clustering the cells and for nonlinear dimension methods mentioned below to identify populations of cells with similar expression profiles. To account for technical variations between samples from different experimental batches, we corrected the PCA embeddings using the Harmony algorithm [4], a method that iteratively clusters and corrects the PC coordinates to adjust for batch specific effects. We assessed the integration of these datasets by employing PCA visualizations of batches of cells and calculating the average silhouette width (ASW) score[5,6] for each cell type population based on the top 20 principal components before correction and the top 20 harmony components after batch correction reported with Wilcoxon rank sum test and p-value significance threshold of 0.05. We clustered cells based on graph-based methods (SNN and Louvain community detection method) using the top 20 Harmony-adjusted components and used the Unifold Manifold Approximation and Projection (UMAP) algorithm[7] to visualize clusters of cells and other known annotations.

Identification and classification of cell types. We used a multi-modal approach to identify immune subpopulations in the NECS dataset. First, we used the 10 cell-surface protein immune cell marker panel of expression to identify main immune cell types. We then further partitioned the main immune cell types into immune subtypes using graph-based clustering and based on the expression of immune cell type signatures from literature [8,9]. The average expression score of each signature was calculated for a single cell by calculating the average scaled expression of all genes within a signature, with the scaling based on the expression of a control set of genes (AddModuleScore function in Seurat[3]), and by taking the absolute value of the average scaled expression to compare scores across signatures within a cell population.

Publicly available datasets:

Data collection, filtering, PCA analysis, and clustering. We downloaded the raw UMI matrix for the scRNA-seq dataset of PBMCs from 45 younger age individuals of European descent [10], which we will refer to as NATGEN. We also downloaded the raw UMI matrix for the scRNA-seq dataset of PBMCS from 7 supercentenarians and 5 younger age individuals of Japanese descent[11], which we will refer to as PNAS. For both PBMC datasets, we performed all downstream processing including filtering, normalization, and scaling of data using the Seurat v.3 [3]. For both datasets, we performed quality control steps based on the number of genes and UMIs detected per cell, and percent of mitochondrial genes expressed per cell. For the NATGEN dataset, we filtered cells based on similar thresholds from the original manuscript[10]. For the PNAS dataset, we filtered cells as previously published[11]. After filtering the datasets, we normalized the expression levels of each cell to compare gene expression between sample cells; gene counts for each cell were normalized by total expression, multiplied by a scale factor of 10,000 and transformed to a log scale. We then performed PCA analysis based on the top 2,000 highly variable genes detected and clustered cells based on graph-based methods (SNN and Louvain community detection method)[3] based on the top significant PCs for each data set implemented in Seurat. We used the UMAP algorithm[4] to visualize the clusters of single cells and other known annotations.

Subpopulation Identification and Harmonization. To define the set of consensus immune cell types across regular aging and longevity, we first identified subpopulations of each cell type using immune cell type signatures from literature [8,9]. The average expression score of each signature was calculated for each single cell across datasets as described for the NECS dataset using Seurat. In addition, we compared the expression of canonical gene markers of the immune populations identified for comparison [12]. After identifying subpopulations in each dataset, we then integrated these scRNA-seq datasets of PBMCs by correcting the PCA embeddings using the Harmony algorthim[4] and assessed batch correction using average silhouette width (ASW) score[5,6] for each cell type population based on the top 20 principal components before correction and the top 20 harmony components after batch correction reported with Wilcoxon statistic and p-value significance threshold of 0.05. We determined the four age groups in the integrated datasets by grouping subjects into four approximate quantile groups based on age in decades.

Statistical Methods.

Heterogeneity of the overall cell type distribution. We compared the overall cell type composition differences across samples and age groups by calculating the cell type diversity statistic for each sample[13]. The cell type diversity statistic E_s is the normalized Shannon entropy based on the cell type proportions p_i for the sample s. The normalization is based on the total number of cell types k to make the entropy measure independent of the number of cell types.

$$E_s = \frac{-\sum_{i=1}^{k} p_{i_s} \, log(p_{i_s})}{log(k)} - 1$$

The normalization is important because we do not have the same number of cell types in the various age groups. For example, if we compare two groups with perfectly uniform cell type proportions, but different number of cell types, the Shannon entropy would show a difference while both distributions are uniformly distributed. We performed ANOVA to assess differences between age groups, and used p-value < 0.05 to determine statistical significance.

Effect of age and sex on cell type distribution. To investigate cell type specific differences across age and sex groups, at subject level, we applied a Bayesian multinomial regression model to the cell type abundances. The outcome of this analysis was the vector of counts of the 13 cell types in each subject, and the exposure was age group and the confounder was sex. Since high throughput sequencing data is compositional [14–16] and cell type frequencies are constrained to the total cells per sample [17], we used a multinomial model that accounts for this constraints and estimates cell type proportions that add up to 1 for each sample. We implemented a Bayesian analysis in the R-package rjags that allows for the fitting of a multinomial regression model without any additional specification. The model set up is

$$Y_{i,1:J} \sim Multinomial(p_{i,1:J}\,, N.total_i)$$
$$log(q_{i,j}) = \alpha_j + \beta_{age.group_i,j} + \gamma_{sex_i,j}$$
$$p_{i,j} = \frac{q_{i,j}}{\sum_{k=1}^{N.ct} q_{i,k}}$$

$$\alpha_j \sim Normal(0, 0.001)$$
$$\beta_{age.group_i,j} \sim Normal(0, 0.001)$$
$$\gamma_{sex_i,j} \sim Normal(0, 0.001)$$

where $Y_{i,1:J}$ represents the abundances of cell type $1{:}J$ for sample i that are modeled using a Multinomial distribution with probabilities $p_{i,1:J}$ $\sum_{j=1}^{J} Y_{i,j} = N.total_i$ for all sample i and $\sum_{j=1}^{J} p_{i,j} = 1$. The probabilities $p_{i,1:J}$ depend on age and sex through the function $log(q_{i,j})$. We chose a reference for each age group (younger age) and sex (male). The implementation of the analysis in rjags does not require a logistic parameterization since rjags can work with unnormalized probabilities that are internally summed up to 1 [18]. This allows us to calculate explicitly the predicted probabilities of all the cell types for every combination of age and sex, with no need for a reference cell type (Supplementary Table S10). Therefore, for each cell type$1{:}j$, we can estimate the probability $p_{i,j}$ for each group profile i from the estimates of the parameters α_j , $\beta_{age.group_i,j}$, $\gamma_{sex_i,j}$ as shown below:

$$\hat{p}_{i,j} = \frac{\exp\,(\hat{\alpha}_j + \hat{\beta}_{age.group_i,j} + \hat{\gamma}_{sex_i,j})}{\sum_{k=1}^{N.ct} \exp\,(\hat{\alpha}_j + \hat{\beta}_{age.group_i,j} + \hat{\gamma}_{sex_i,j})}$$

where the notation ^ represents the estimated parameters using rjags.

The model was estimated using Markov Chain Monte Carlo (MCMC) sampling using rjags, the R package for JAGS[19]. We ran parameter inference for all coefficients for group level predicted probabilities of composition ($p_{i,j}$) and the age group and sex effect coefficients (β, γ) using 1,000 iterations with 500 iterations for burn-in. We estimated the group level composition predicted probabilities and 95 percent credible interval for males and female subjects for younger, middle, older, and EL age groups across all immune cell types. In addition, to assess the significance of the effect of age and sex in each cell type, we calculated the z-score (Z) for parameters β and γ based on the mean estimate and standard error of the posterior distribution. We subsequently calculated the two-sided large sample p-value based on the standard normal distribution: $2\Phi(-|Z|))$ where Φ is the standard normal cumulative distribution function. We calculated the adjusted p-value based on the Benjamin and Hochberg correction for multiple testing across all coefficients tested.

In addition, to investigate cell type specific differences in other age group comparisons including older age vs. EL, we applied the multinomial regression analysis as described above with the exception of the age group reference set to EL.

Analysis of the hierarchy of peripheral immune compartments. To estimate the hierarchy of peripheral immune compartments, we utilized K2Taxonomer (v1.0.5)[20], which performs top-down partitioning of cell types based on the relative similarity of their transcriptomic profiles. Prior to running K2Taxonomer, we performed several data processing steps. First, for each subject, we removed profiles of lowly represented cell types with fewer than 10 profiles for that subject only. Following this filter, we further removed all plasma cell profiles because they were represented in fewer than 10 subjects. Next, for each subject we aggregated single-cell profiles of each cell type into a “pseudo-bulk” profile by summing the counts of each gene, followed by normalization to log2(counts-per-million). The resulting data set included 515 total profiles across 66 subjects and 12 cell types. The number of subjects for which each of these 12 cell types were identified in each of the four batches is given in Supplementary Table S12. Next, we removed lowly expressed genes, i.e., genes that failed to reach 2 counts-per-million in at least 2 profiles across all batches, which left 12,354 genes. Finally, we performed batch correction on these data using ComBat (v3.40.0)[21], parameterized to preserve the cell type-specific signals using “mod” parameter of the ComBat() function. K2Taxonomer was run on these data, using the package’s “group-level” workflow, and setting the number of features parameter, "nFeats", to use 5% of the total genes, i.e., 618 genes per partition estimate. Finally, we calculated the cell type diversity statistics for each K2Taxonomer generated cell type subgroup.

Cell type specific differential gene expression analysis. To investigate the cell type specific differences across age groups, we used a Bayesian mixed effects model with the rjags R package[19] to perform differential gene expression analysis across the four age groups. For each gene, we applied the model to the expression values across cells from all subjects with exposure variable for age group comparing middle, older, and EL age groups in reference to the younger age group, and confounding variables including sex, ethnicity, and batch, and a random effect that accounts for within subject correlations of cells. For each cell type, we first filtered genes to keep genes with expression in at least fifty percent of the smallest cell type population. We then applied the Bayesian mixed effects model where for each cell type, to each gene i:

$$Gene_i \sim Normal(\mu_i \,, \tau)$$

$$\mu_i \;=\; \beta_{S,sample} \;+\; \beta_1 Middle_i \;+\; \beta_2 Older_i \;+\; \beta_3 EL_i \;+\; \beta_4 Sex_i \;+\; \beta_5 Ethnicity_i \;+\; \beta_6 Batch_i$$

$$\beta_{S,sample} \sim Normal(\beta_0 \,, \tau_s)$$
$$\beta_k \sim Normal(0, 0.0001) \,\forall\, k \;\in\; [0, \ldots, 6]$$
$$\tau \sim Gamma(0.0001, 0.0001)$$
$$\tau_s \;\sim Gamma(0.00001, 0.00001)$$

where $Gene_i$ is the log-normalized expression of a gene for each cell i in the cell type; $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_0$ are model parameters. The subject specific intercept $\beta_{S,sample}$ follows a normal distribution with population mean $, \beta_0$. The model is adjusted by fixed covariates sex, batch, and ethnicity, as well as a random effect based on samples to account for differences in cell abundances between samples within groups. Using this model, we monitored the age-dependent coefficients (β_1, β_2, β_3) across 10,000 MCMC iterations with 2,500 burn-in iterations to obtain the log fold change (logFC) based on age group. Then, we calculated the z-score (Z) for the age-dependent parameter based on the mean estimate and standard error of the posterior distribution. We subsequently calculated the two-sided p-value based on the standard normal distribution: 2*Φ(-|Z|)) where Φ is the standard normal cumulative distribution function. We calculated the FDR based on the Benjamin and Hochberg correction for multiple testing across all genes tested. Significant differential genes were selected based on a significance of FDR < 0.05 and fold change cutoff of 10 percent (|logFC| > log(1.1)).

In addition, to investigate cell type specific differences between other age group comparisons including older age vs. EL, we applied the mixed effects model as described above with the exception of the age group reference set to EL.

Bulk level differential gene expression analysis. We performed differential gene expression analysis at the bulk level between age groups using DESeq2 [22]. We filtered genes in the single cell data to keep genes expressed in at least 50% of the smallest cell type population. We then aggregated the raw counts per sample and ran DESeq2 to perform normalization and fit a negative binomial generalized linear model with covariates age group, sex, batch, and ethnicity. We used Wald test to identify differentially expressed genes between middle, older, and EL v. younger age with a log fold change greater than log2(1.5) and FDR < 0.05.

References

1 Park LM, Lannigan J, Jaimes MC. OMIP-069: Forty-Color Full Spectrum Flow Cytometry Panel for Deep Immunophenotyping of Major Cell Subsets in Human Peripheral Blood. *Cytometry Part A* 2020; **97**: 1044–51.

2 Alpert A, Pickman Y, Leipold M, *et al.* A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nature Medicine* 2019; **25**: 487–95.

3 Stuart T, Butler A, Hoffman P, *et al.* Comprehensive Integration of Single-Cell Data. *Cell* 2019; **177**: 1888-1902.e21.

4 Korsunsky I, Millard N, Fan J, *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods* 2019; **16**: 1289–96.

5 Tran HTN, Ang KS, Chevrier M, *et al.* A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biology* 2020; **21**: 12.

6 Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 1987; **20**: 53–65.

7 McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv:180203426 [cs, stat]* 2020; published online Sept 17. http://arxiv.org/abs/1802.03426 (accessed April 12, 2021).

8 Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling tumor infiltrating immune cells with CIBERSORT. *Methods Mol Biol* 2018; **1711**: 243–59.

9 Popescu D-M, Botting RA, Stephenson E, *et al.* Decoding human fetal liver haematopoiesis. *Nature* 2019; **574**: 365–71.

10 van der Wijst MGP, Brugge H, de Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nature Genetics* 2018; **50**: 493–7.

11 Hashimoto K, Kouno T, Ikawa T, *et al.* Single-cell transcriptomics reveals expansion of cytotoxic CD4 T cells in supercentenarians. *PNAS* 2019; **116**: 24242–51.

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

13 Karagiannis TT, Monti S, Sebastiani P. Cell Type Diversity Statistic: An Entropy-Based Metric to Compare Overall Cell Type Composition Across Samples. *Frontiers in Genetics* 2022; **13**. https://www.frontiersin.org/article/10.3389/fgene.2022.855076 (accessed April 10, 2022).

14 Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology* 2016; **26**: 322–9.

15 Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* 2017; **8**. DOI:10.3389/fmicb.2017.02224.

16 Lin H, Peddada SD. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *npj Biofilms Microbiomes* 2020; **6**: 1–13.

17 Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* 2019; **15**: e8746.

18 Lunn D, Jackson C, Best N, Thomas A, Spiegelhalter D. The BUGS Book: A Practical Introduction to Bayesian Analysis. London: Chapman Hall, 2013.

19 Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008; **9**: 523–39.

20 Reed ER, Monti S. Multi-resolution characterization of molecular taxonomies in bulk and single-cell transcriptomics data. *Nucleic Acids Res* 2021; **49**: e98.

21 Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007; **8**: 118–27.

22 Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 2014; **15**: 550.