nature portfolio

Peer Review File

Sex-specific trajectories of nonlinear immune aging at single cell level

Corresponding Author: Professor Jacques Behmoaras

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The manuscript entitled “Sex-specific trajectories of nonlinear immune aging at single cell level” by Park et al. investigates sex-dependent dynamics of immune aging across the human lifespan using large-scale single-cell transcriptomic data. Using a multi-cohort integration of peripheral blood mononuclear cell (PBMC) scRNA-seq datasets and pseudobulk analytical frameworks, the authors report nonlinear trajectories of immune aging with distinct transcriptional inflection points around midlife (~40 years) and late life (>60 years), accompanied by sex-specific differences in particular functional pathways. The study is interesting and technically generally sound, but I have several major concerns that need to be addressed.

1) Authors integrated and analyzed four large-scale scRNA-seq datasets (line 68) — they likely encountered dataset shift (batch effects). How exactly was this addressed? I would like to see a combined UMAP colored by dataset, which would demonstrate how well the datasets are mixed after integration. Currently, there is no clear description of dataset integration in the Methods. It almost appears that the datasets may have been analyzed separately — for example, pseudobulk values could have been calculated independently per dataset. This is not clearly stated. If this is the case, I am not convinced it is appropriate to state in the Results that “We integrated and analyzed four large-scale scRNA-seq datasets” (line 68), as the term “integrated” implies joint embedding or harmonization at the single-cell level. In the Methods, the authors instead state that “Pseudobulk expression matrices from all datasets were merged” (line 721). It is unclear what “merged” means in this context — were values averaged, concatenated, or simply combined across individuals? A more precise description is required. Moreover, the issue of batch effects remains: even at the pseudobulk level, the authors should demonstrate that dataset-specific biases are adequately controlled.

2) The authors mention that “The merged pseudobulk expression matrix was corrected for dataset-specific batch effects using removeBatchEffect function from limma” (line 740), but this statement appears after the “Sex-stratified differential expression sliding window analysis (DE-SWAN)” section (line 723) in the Methods. Does this imply that DE-SWAN was performed on uncorrected data? If so, this would be a significant methodological flaw, as batch effects could strongly bias differential expression results.

3) In any case, I would like to see diagnostic evidence of batch effect correction at the pseudobulk level — for example, a PCA plot of samples colored by dataset before and after correction, demonstrating that samples do not cluster by dataset after correction. This is standard practice and necessary for validating batch effect correction quality.

4) The authors performed cell type proportion analyses (line 72) based on scRNA-seq data — however, such estimates are known to be biased due to technical artifacts including differential capture efficiency and cell-type-specific dissociation sensitivity. These issues are well recognized in the field, and dedicated computational frameworks have been developed to mitigate them (e.g., see https://doi.org/10.1093/bioinformatics/btac582). Could the authors validate their findings using an orthogonal method (e.g., flow cytometry or deconvolution of bulk RNA-seq), or at least discuss these limitations? At a minimum, since multiple independent datasets are available, the authors should treat them as biological replicates and demonstrate that inferred cell type proportions are consistent across datasets.

5) The authors generated pseudobulk expression profiles for five major PBMC cell types: CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes (line 79). However, proportions of subtypes within these major populations (e.g., naive vs memory

T cells) are known to change substantially with age and could confound the observed transcriptional changes. Could the authors demonstrate that subtype composition remains relatively stable across age, or alternatively perform analyses at a finer cell subtype resolution?

6) The authors report two peaks of differential expression — around ~40 years and after 60 years (line 89). These peaks appear to coincide with higher sample numbers in these age ranges. Could this observation be partially driven by increased statistical power rather than true biological effects? The authors should control for sample size effects, for example by downsampling or applying methods that account for uneven age distribution.

7) Which gene background was used for functional enrichment analyses? No information is provided in the Methods (lines 769–778), suggesting that the default background may have been used. This is problematic: enrichment analyses should use expressed genes within each cell type as background. Otherwise, results may largely reflect general cell-type-specific functions rather than trajectory-specific biological processes.

8) Given that four independent datasets were used, it would be more appropriate to train and test the aging clock models across datasets (e.g., leave-one-dataset-out validation) rather than relying solely on five-fold cross-validation (line 817). This would provide a more robust assessment of generalizability and reduce the risk of overfitting.

9) Finally, aging clocks trained to predict chronological age have been increasingly criticized in the aging research community (e.g., see https://doi.org/10.1038/s43587-026-01066-6, https://doi.org/10.1038/s41467-025-66106-y). Second-generation clocks that predict clinically relevant outcomes such as mortality or multimorbidity are now generally preferred. While such outcome data may not be available in this study, the authors should explicitly acknowledge the limitations of first-generation clocks and discuss how this affects interpretation of their results.

Dr. Ekaterina Khrameeva
Associate Professor
Skolkovo Institute of Science and Technology

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this study, Park et al. investigate the transcriptional changes in PBMCs in men and women during aging, using publicly available single cell RNA-seq from over 1800 individuals between 19 and 97 years old. The authors report distinct peaks of differential gene expression in different cell types (CD4 vs CD8) at different ages (40 vs 60), as well as sex-specific differences such as sex-dependent immune cell activation. They then apply machine learning based aging clocks to show that sex-stratified models outperform sex-combined models in learning from specific immune trajectories.

The manuscript is well written and the findings are clearly described. While sex differences in immune aging are known and it is not surprising that aging clocks perform better on sex-stratified data, there is still a need for in depth understanding of the sex differences and how they contribute to sex dimorphisms in disease. However, in the enclosed study, the robustness of the findings is difficult to assess due to lack of quality control information. In addition, the claims related to senescence are not supported. These points are described below.

A major concern with the study is the uneven distribution of ages across the 4 public datasets. Most samples come from Onek1k and AIDA, which are enriched for an average age of ~40 (AIDA) and ~65 (Onek1k), which are the ages that appear to have the most DEFs. Further, since the individual datasets are clearly capturing either the first or second peak, it is plausible that the bimodal nature of the result is simply due to the uneven age distribution in the combined dataset.
Related to the above point, although the authors include batch correction in the methods, there are no data (e.g. feature plots) showing batches before and after correction. This is important quality control, as it is possible that the different DE signatures in the two peaks are driven by differences in the AIDA and Onek1k datasets themselves.

Data on the composition of each dataset should also be shown as this could bias the cell type analysis.

The claim that the female-specific late-increase cluster represents a senescence-associated cellular state is not supported. Senescent cells are rare, even in aged individuals and their gene expression signatures are well defined. There is no actual analysis of senescence in this study and this claim (included in the abstract) is purely speculative.

Minor:
Extended Fig. 4 is not readable.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have sufficiently addressed my comments and concerns, and the manuscript has been significantly improved. I have only one minor suggestion: please ensure that all additional analyses performed in response to the reviewers' comments are incorporated into the revised manuscript. For example, Figure R2 does not appear to be included. Although these analyses may seem technical, including them is important for demonstrating the robustness of the presented results and will strengthen the manuscript.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
My concerns have been address and I am supportive of publication.

(Remarks on code availability)

Reviewer #1 (General comments)

The manuscript entitled “Sex-specific trajectories of nonlinear immune aging at single cell level” by Park et al. investigates sex-dependent dynamics of immune aging across the human lifespan using large-scale single-cell transcriptomic data. Using a multi-cohort integration of peripheral blood mononuclear cell (PBMC) scRNA-seq datasets and pseudobulk analytical frameworks, the authors report nonlinear trajectories of immune aging with distinct transcriptional inflection points around midlife (~40 years) and late life (>60 years), accompanied by sex-specific differences in particular functional pathways. The study is interesting and technically generally sound, but I have several major concerns that need to be addressed.

Response

We sincerely thank the Reviewer #1 for reading carefully the manuscript and providing highly constructive feedback. We have taken into account all of Reviewer #1’s suggestions and changed the manuscript which improved considerably.

Reviewer #1; Comments 1 to 3 (Batch effect correction)

1) Authors integrated and analyzed four large-scale scRNA-seq datasets (line 68) — they likely encountered dataset shift (batch effects). How exactly was this addressed? I would like to see a combined UMAP colored by dataset, which would demonstrate how well the datasets are mixed after integration. Currently, there is no clear description of dataset integration in the Methods. It almost appears that the datasets may have been analyzed separately — for example, pseudobulk values could have been calculated independently per dataset. This is not clearly stated. If this is the case, I am not convinced it is appropriate to state in the Results that “We integrated and analyzed four large-scale scRNA-seq datasets” (line 68), as the term “integrated” implies joint embedding or harmonization at the single-cell level. In the Methods, the authors instead state that “Pseudobulk expression matrices from all datasets were merged” (line 721). It is unclear what “merged” means in this context — were values averaged, concatenated, or simply combined across individuals? A more precise description is required. Moreover, the issue of batch effects remains: even at the pseudobulk level, the authors should demonstrate that dataset-specific biases are adequately controlled.

2) The authors mention that “The merged pseudobulk expression matrix was corrected for dataset-specific batch effects using removeBatchEffect function from limma” (line 740), but this statement appears after the “Sex-stratified differential expression sliding window analysis (DE-SWAN)” section (line 723) in the Methods. Does this imply that DE-SWAN was performed on uncorrected data? If so, this would be a significant methodological flaw, as batch effects could strongly bias differential expression results.

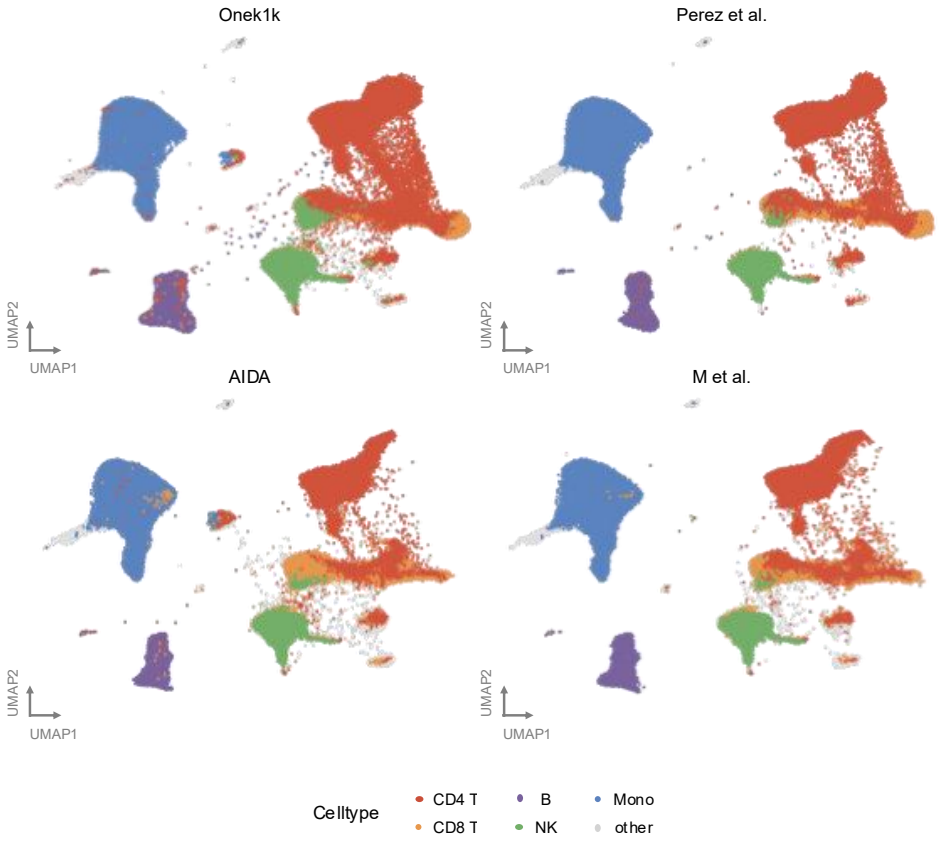
3) In any case, I would like to see diagnostic evidence of batch effect correction at the pseudobulk level — for example, a PCA plot of samples colored by dataset before and after correction, demonstrating that samples do not cluster by dataset after correction. This is standard practice and necessary for validating batch effect correction quality.

Response

We thank the Reviewer for raising these important points regarding dataset integration and batch effect correction. We agree with the Reviewer that these steps were insufficiently described in the original manuscript and have now clarified both the analytical workflow (lines 606-607, 619, 627-628, 644-647) and the terminology used throughout the manuscript (lines 19, 48, 52-53, 68, 76, 99, 294, 375, 389, 396, 1077-1080, 1159-1161).

The Reviewer is correct that the term “integrated” incorrectly implies joint embedding or harmonization at the single-cell level and we have revised the manuscript accordingly to avoid this interpretation. To further clarify, our study did not perform single-cell-level integration or harmonization using methods such as Seurat integration, Harmony, or scVI. Instead, each dataset was independently projected onto the Azimuth human PBMC reference to obtain consistent cell type annotations across datasets. Based on these annotations, donor-level pseudobulk expression matrices were generated separately for each dataset and subsequently combined by concatenating matched gene features across donors for downstream analyses.

In line with the Reviewer’s suggestion, we demonstrate the consistency of cell type annotation across all datasets through a UMAP plot colored by major immune cell types and split by dataset (Supplementary Figure 1 in the revised manuscript). This visualization shows that corresponding immune cell populations from the four datasets map consistently onto the PBMC reference space.



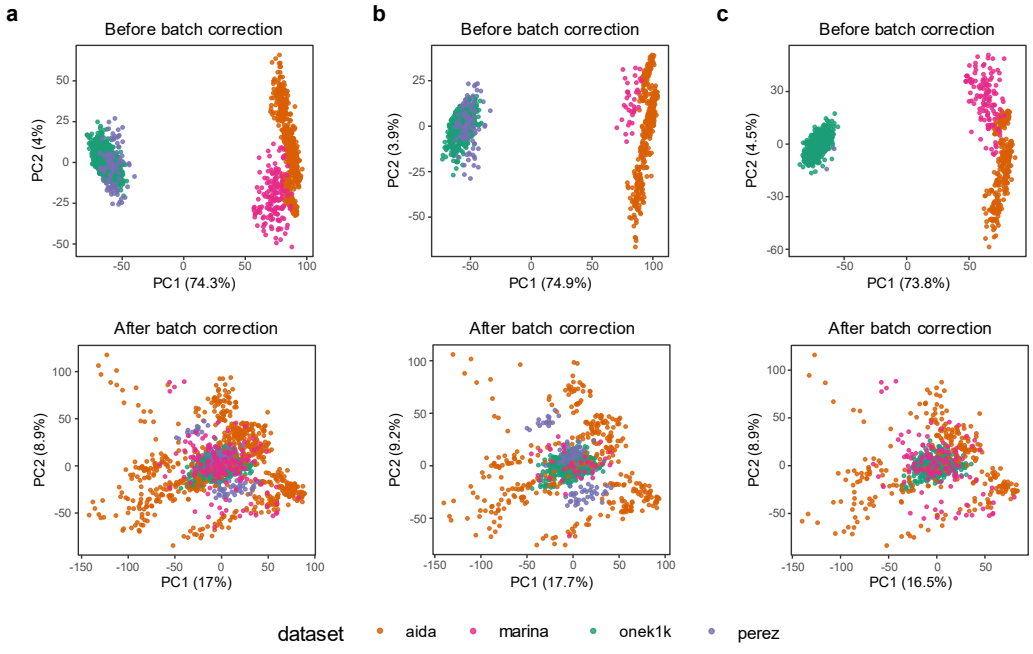
Supplementary Figure 1. UMAP projection of major PBMC cell types across the four datasets.

Regarding batch effects, we accounted for dataset-specific variation separately in the two main downstream analyses.

Importantly, we would like to clarify that DE-SWAN was not performed on uncorrected data. Dataset identity was included as a covariate in the linear modeling framework used for differential expression testing. Batch-associated variation was therefore controlled directly within the regression model during differential expression analysis. According to the Reviewer's suggestion, we have revised the Methods section to clarify this point more explicitly (lines 627-628).

Second, for the downstream clustering analysis of significant age-associated features (SAFs) identified by DE-SWAN, we applied limma's removeBatchEffect function to the merged pseudobulk expression matrices (performed separately for both-sex, female, and male analyses) prior to trajectory clustering. In this context, batch correction was applied specifically for clustering and visualization purposes.

As requested by the Reviewer, we now include PCA plots before and after batch effect correction at the pseudobulk level (Supplementary Figure 5 in the revised manuscript). Prior to correction, samples exhibited partial separation by dataset, whereas this dataset-associated structure was substantially reduced after correction, indicating effective mitigation of batch effects.



Supplementary Figure 5. Batch effect correction for pseudobulk expression matrices. **a-c**, PCA plots before and after batch effect correction for both-sex (**a**), female (**b**), and male (**c**) pseudobulk expression matrices.

We additionally evaluated whether applying removeBatchEffect prior to DE-SWAN altered the results. The overall findings remained highly consistent with the original analysis (Figure R1), including the presence of two major peaks in both sexes, larger peak magnitudes in females, enrichment of CD4 T cell features in the first peak around 40 years of age, and increased contribution of CD8 T cell features in the second peak after approximately 60 years of age.

These results demonstrate that primary conclusions are robust to alternative batch correction strategies.

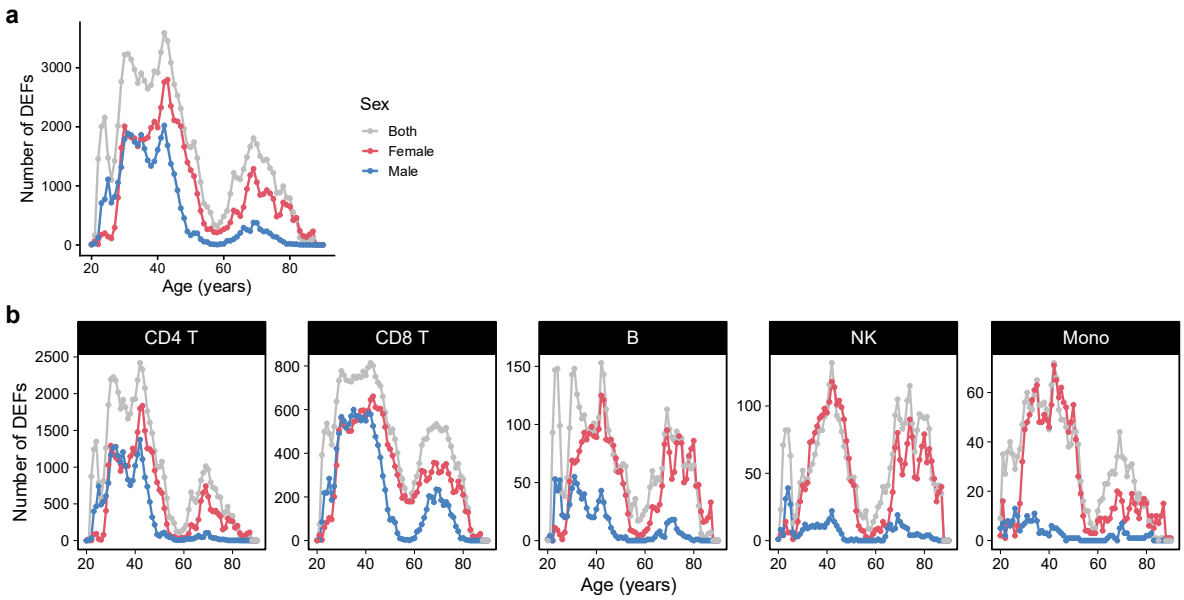


Figure R1. DE-SWAN results after applying removeBatchEffect. **a**-**b**, Number of differentially expressed features (DEFs, q < 0.05) across the age range identified by DE-SWAN algorithm after applying limma's removeBatchEffect, shown for all major PBMC cell types combined (**a**) and for each cell type separately (**b**).

Reviewer #1; Comment 4

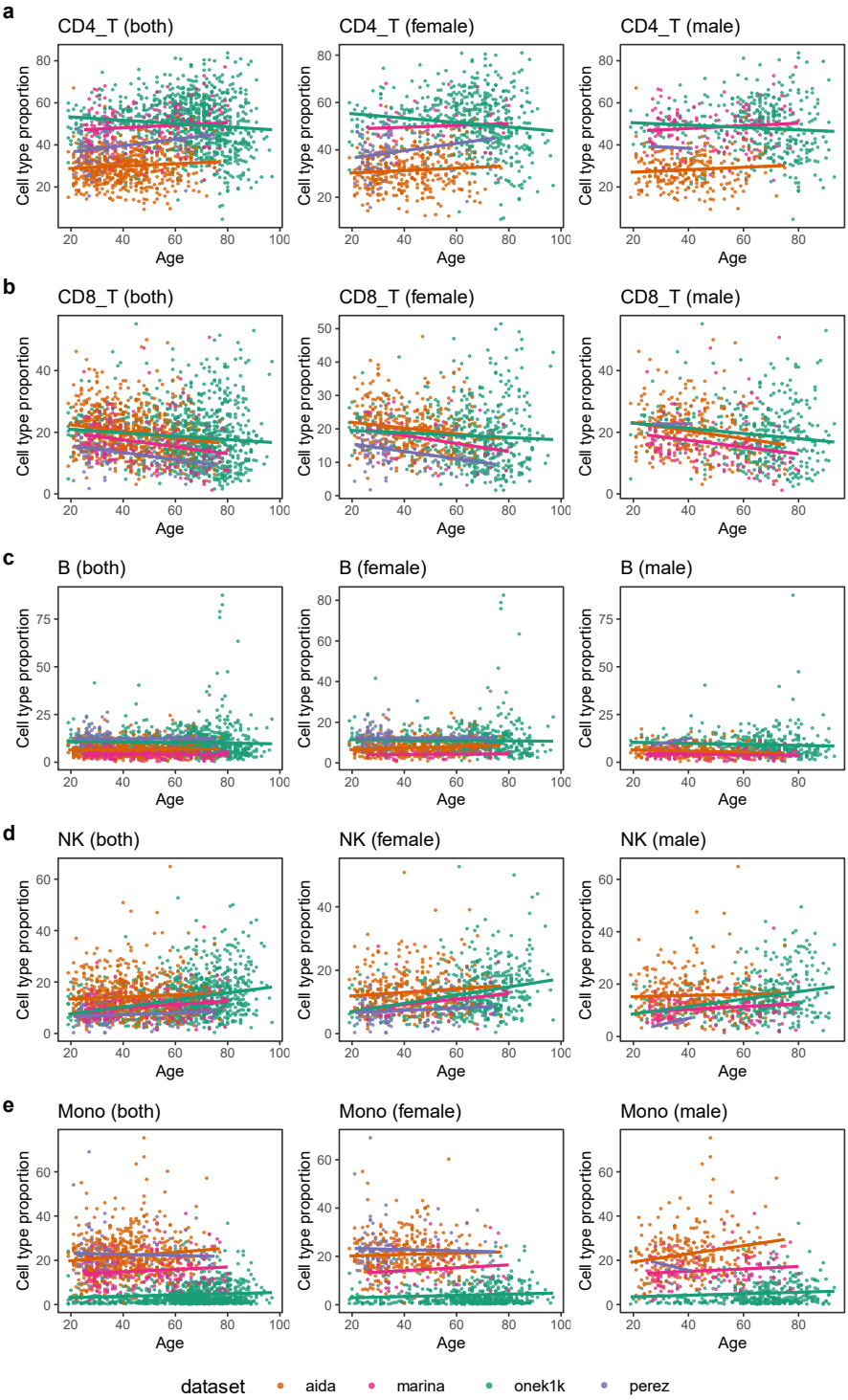
4) The authors performed cell type proportion analyses (line 72) based on scRNA-seq data — however, such estimates are known to be biased due to technical artifacts including differential capture efficiency and cell-type-specific dissociation sensitivity. These issues are well recognized in the field, and dedicated computational frameworks have been developed to mitigate them (e.g., see https://doi.org/10.1093/bioinformatics/btac582). Could the authors validate their findings using an orthogonal method (e.g., flow cytometry or deconvolution of bulk RNA-seq), or at least discuss these limitations? At a minimum, since multiple independent datasets are available, the authors should treat them as biological replicates and demonstrate that inferred cell type proportions are consistent across datasets.

Response

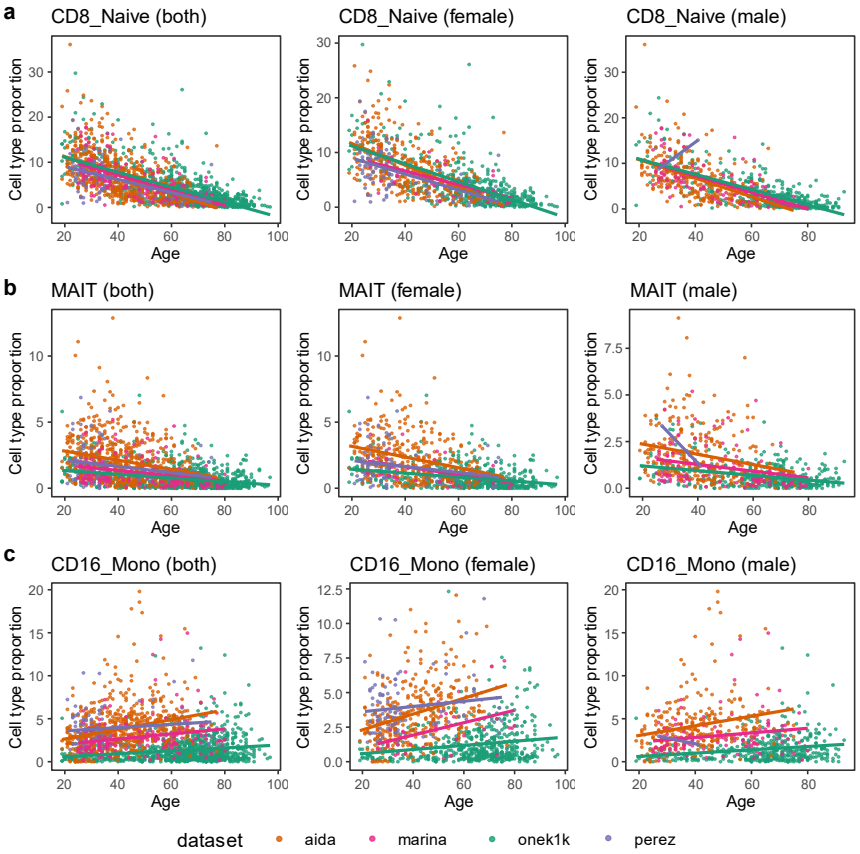
We agree with the Reviewer's assessment on the potential technical biases in scRNA-seq-based cell type proportion analyses. The Reviewer is correct that inferred cell type proportions can be influenced by dataset-specific factors, including differences in cell capture efficiency, dissociation sensitivity, and sample processing.

As suggested by the Reviewer, we revised the cell type proportion analysis using the propeller framework (speckle R package). Propeller is specifically designed for differential cell type proportion analysis while accounting for biological replication and experimental covariates. Instead of assessing simple correlations between age and cell type proportions in the merged dataset, we modelled donor-level cell type proportions using a linear framework with dataset identity included as a covariate. The updated results are presented in the revised Table 1 (line 1255), and the relevant Results and Methods section has been revised accordingly (lines 72-75, lines 398-401, lines 689-699).

In addition, to evaluate the consistency of cell type proportion estimates across datasets, we generated dataset-stratified visualizations of cell type proportions across age. Supplementary Figures 2–3 in the revised manuscript show the proportions of the five major cell types (CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes), as well as the cell subtypes highlighted in the manuscript (lines 398-401; CD8 naïve T cells, MAIT cells, and CD16 monocytes). Overall, similar age-associated trends were observed across the independent cohorts, supporting that the findings are not driven by a single dataset.



Supplementary Figure 2. Age-associated changes in major cell type proportions across the four datasets. **a**-**e**, Cell type proportions plotted against age for CD4 T cells (**a**), CD8 T cells (**b**), B cells (**c**), NK cells (**d**), and monocytes (**e**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.



Supplementary Figure 3: Age-associated changes in detailed cell subtype proportions across the four datasets. **a**-**c**, Cell type proportions plotted against age for CD8 Naïve T cells (**a**), MAIT cells (**b**), and CD16 monocytes (**c**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.

To follow up on the Reviewer's suggestion, we have now explicitly discussed the limitations of scRNA-seq-based cell type proportion inference and the potential influence of technical variability in the Discussion section (lines 490-493).

Reviewer #1; Comment 5

5) The authors generated pseudobulk expression profiles for five major PBMC cell types: CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes (line 79). However, proportions of subtypes within these major populations (e.g., naive vs memory T cells) are known to change substantially with age and could confound the observed transcriptional changes. Could the authors demonstrate that subtype composition remains relatively stable across age, or alternatively perform analyses at a finer cell subtype resolution?

Response

This is indeed a crucial point. We agree with the Reviewer that substantial shifts in immune subtypes occur with aging, particularly within T cell compartments, where naïve CD4 and CD8 T cells decrease while memory and effector populations increase. Consistent with this, our updated cell type proportion analysis (Table 1) also shows significant age-associated changes in these subpopulations. This raises the possibility that transcriptional changes observed in aggregated CD4 and CD8 T cell populations could be driven by shifts in subtype composition.

To address the Reviewer's point, we performed DE-SWAN analysis at a finer cellular resolution for CD4 and CD8 T cells that exhibited strong signals in the original analysis (Figure R2). Specifically, we analyzed CD4 naïve T cells (CD4 Naïve), CD4 central memory T cells (CD4 TCM), CD4 effector memory T cells (CD4 TEM), CD8 naïve T cells (CD8 Naive), CD8 central memory T cells (CD8 TCM), and CD8 effector memory T cells (CD8 TEM).

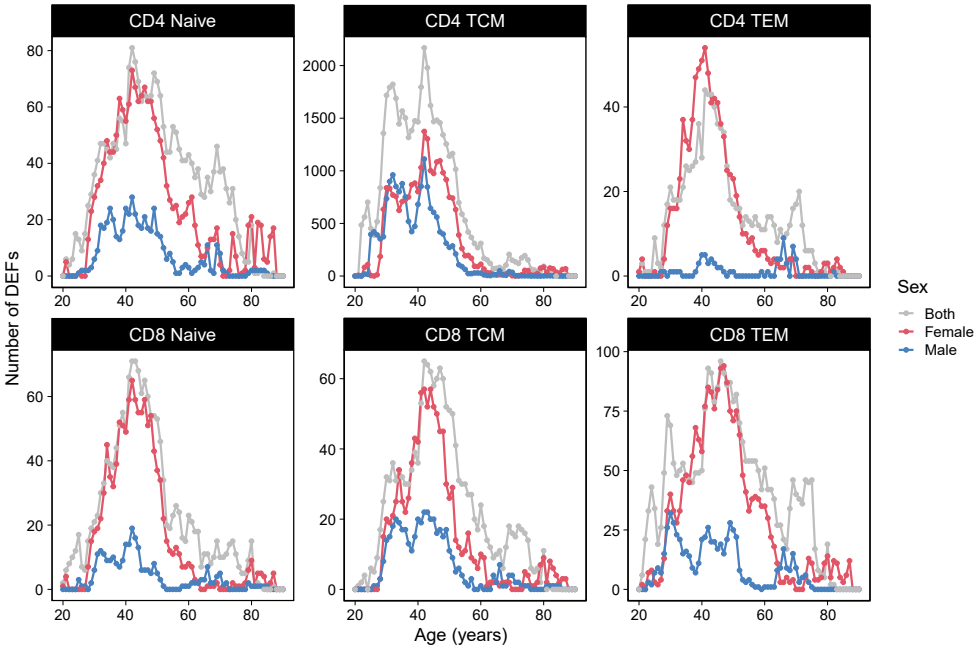


Figure R2. DE-SWAN results of CD4 and CD8 T cell subtypes. Number of differentially expressed features (DEFs, q < 0.05) across the age range identified by DE-SWAN algorithm in CD4 naïve T cells (CD4 Naïve), CD4 central memory T cells (CD4 TCM), CD4 effector memory T cells (CD4 TEM), CD8 naïve T cells (CD8 Naive), CD8 central memory T cells (CD8 TCM), and CD8 effector memory T cells (CD8 TEM).

Notably, the results show that the major transcriptional aging patterns observed in aggregated CD4 and CD8 T-cell populations are largely preserved at the subtype level. In particular, a prominent peak of differentially expressed features around ~40 years of age is consistently observed across multiple T-cell subtypes, with a second, weaker peak after ~60 years also detectable in several subtypes, particularly in the both-sex analysis (Figure R2).

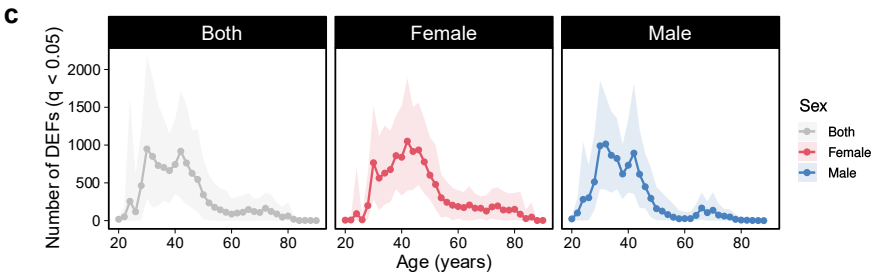
Together, these results indicate that the transcriptional aging signatures identified in aggregated CD4 and CD8 T-cell populations are not solely driven by shifts in subtype composition. Rather, similar nonlinear aging patterns are detectable within individual T-cell subtypes, supporting the robustness of the observed age-associated transcriptional trajectories.

Reviewer #1; Comment 6

6) The authors report two peaks of differential expression — around ~40 years and after 60 years (line 89). These peaks appear to coincide with higher sample numbers in these age ranges. Could this observation be partially driven by increased statistical power rather than true biological effects? The authors should control for sample size effects, for example by downsampling or applying methods that account for uneven age distribution.

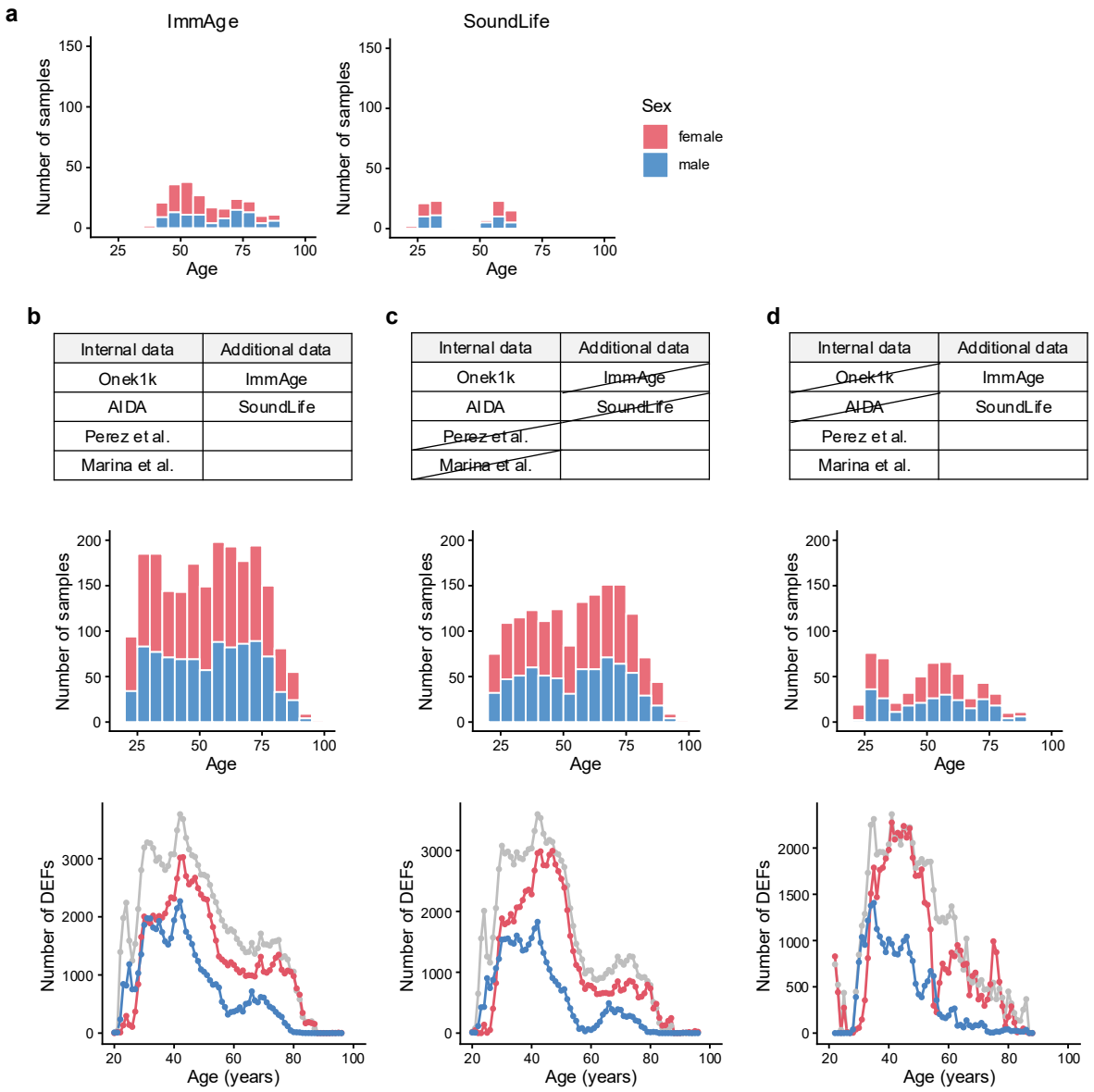
Response

The Reviewer is correct to suggest that the peaks of differential expression may be driven by sample distribution. To address the Reviewer's concern, we performed a modified DE-SWAN analysis with downsampling as suggested by the Reviewer and added the results as Extended Data Figure 2c in the revised manuscript. Specifically, for each age window tested (20-90 years, in 2-year increments), we randomly downsampled to 200 samples and repeated the analysis for 100 iterations. The two peaks remained after downsampling, including the prominent first peak around 40 years of age and a smaller second peak around 65 years of age (most clearly observed in the male analysis). We have added the corresponding methods and results descriptions to the manuscript (lines 637-641, 91-92).



Extended Data Figure 2. DE-SWAN robustness analysis across parameters and cell types. **c**, Number of differentially expressed features (DEFs) across age identified by DE-SWAN using 100 random downsampling iterations, each using 200 samples.

We further evaluated whether the observed bimodal pattern could be driven by the age distributions of the major contributing datasets, AIDA and OneK1K, as discussed in our response to Reviewer #2, Comment #1. Briefly, we incorporated two additional recently published PBMC datasets (ImmAge and SoundLife) and repeated the DE-SWAN analysis using different combinations of datasets (Supplementary Figure 4 in the revised manuscript). Importantly, the bimodal pattern remained evident even when AIDA and OneK1K were completely excluded from the analysis, with peaks observed at similar age ranges. Together, the downsampling analysis and the reproducibility of the two peaks across independent dataset combinations support the conclusion that these aging transitions are unlikely to be explained solely by sample size differences or uneven age distributions and instead reflect underlying biological aging processes.



Supplementary Figure 4. Reproducibility of DE-SWAN peaks using additional PBMC scRNA-seq datasets. **a**, Age and sex distribution of the ImmAge and SoundLife datasets, which were used as additional PBMC scRNA-seq data for validation. **b-d**, Summary of included datasets, their age and sex distributions, and DE-SWAN results under different dataset combinations: all internal and additional datasets (**b**); Onek1k and AIDA datasets only (**c**); and Perez et al., Marina et al., and additional datasets (**d**).

Reviewer #1; Comment 7

7) Which gene background was used for functional enrichment analyses? No information is provided in the Methods (lines 769–778), suggesting that the default background may have been used. This is problematic: enrichment analyses should use expressed genes within each cell type as background. Otherwise, results may largely reflect general cell-type-specific functions rather than trajectory-specific biological processes.

Response

In the original analysis, we used all genes represented in the pseudobulk expression matrix as the background set. This matrix was constructed using features (cell type-gene pairs; cell type-specific expression) that were expressed across all donors in all four datasets (>0 pseudobulk expression in all donors) as mentioned in the Methods section (lines 619-621). While this ensures that only consistently detected genes were included, we agree with the Reviewer that using a cell type-agnostic background is suboptimal because it does not account for cell-type-specific gene expression.

To address this point, we have revised the pathway enrichment analysis using cell-type-specific background gene sets, defined as genes (features) expressed in each cell type across all donors. This approach better accounts for differences in baseline gene expression between cell types and reduces bias toward general cell-type-specific functions.

We have updated the Methods part accordingly (lines 676-678) and revised all affected figures (Fig. 3h, Fig. 3j, Fig. 3m, Extended Data Fig. 4, Extended Data Fig. 5a, Extended Data Fig. 5c, Extended Data Fig. 5g) and results sections (lines 153-157, 165-170, 176-261) to reflect the corrected analysis. Importantly, the revised analysis yielded consistent biological interpretations, with no change to the main conclusions of the functional enrichment results.

Reviewer #1; Comment 8

8) Given that four independent datasets were used, it would be more appropriate to train and test the aging clock models across datasets (e.g., leave-one-dataset-out validation) rather than relying solely on five-fold cross-validation (line 817). This would provide a more robust assessment of generalizability and reduce the risk of overfitting.

Response

We thank the reviewer for this important suggestion. We agree that demonstrating generalizability beyond the discovery datasets is a critical consideration for aging clock models.

To address this, we evaluated the final models on two additional independent PBMC datasets in the revised manuscript (ImmAge and SoundLife; lines 586-604, line 340), which were not used in model training, model selection, or parameter tuning. These datasets therefore provide a further direct assessment of external generalizability. In the ImmAge cohort, correlations were 0.69 (both-sex), 0.78 (female), and 0.80 (male). In the SoundLife cohort, correlations were 0.80 (both-sex), 0.78 (female), and 0.65 (male). These results extend the external validation already presented in the original manuscript and support the robustness and generalizability of the proposed models across additional unseen populations.

In addition, we performed leave-one-dataset-out (LODO) validation as a sensitivity analysis within the discovery cohorts. However, interpretation of LODO results in this setting is challenging because the discovery datasets differ substantially in sample size, age distribution, and sex composition. As a result, holding out a single dataset produces training and testing sets with markedly different demographic structures, making model performance highly dependent on the excluded cohort. This consideration was important for our analysis, as we aimed to (i) compare and select appropriate algorithms (elastic net regression, XGBoost, and MLP) and (ii) evaluate performance differences between sex-combined and sex-stratified models.

To illustrate this, we performed LODO validation using MLP models. Because of strong sex imbalance between Perez et al. (97.9% female) and Marina et al. (77.6% male) (Extended Data Fig. 1b), these two datasets were combined into a single cohort for this analysis, resulting in three effective datasets: Onek1k, AIDA, and the combined Perez/Marina cohort.

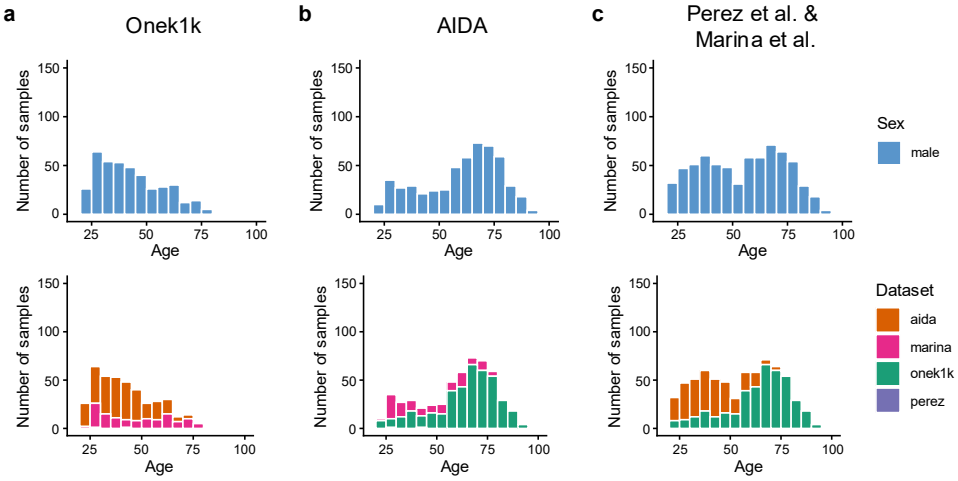


Figure R3. Age and dataset composition of leave-one-dataset-out (LODO) training sets for male MLP models. **a-c**, Age distribution and dataset composition of the LODO training sets when the Onek1k dataset was held out (**a**), the AIDA dataset was held-out (**b**), and the Perez et al. and Marina et al. datasets were held out (**c**).

Held-out dataset	Onek1k	AIDA	Perez et al. & Marina et al.
Number of training samples	879	1206	1571
Mean age of training samples	43.06	60.18	54.81
Number of testing samples	949	622	257
Mean age of testing samples	64.14	40.98	47.54
Internal testing RMSE	29.88	17.16	12.95
External testing RMSE	22.06	16.88	15.86

Table R1. Sample distribution and model performance in leave-one-dataset-out (LODO) training sets for male MLP models.

As shown in the LODO results, the composition of the training set varied substantially depending on the held-out dataset (Figure R3, Table R1). For example, when Onek1k was excluded, the training set contained a smaller sample size with a narrower and younger age distribution, whereas exclusion of AIDA cohort resulted in a markedly older training distribution. In contrast, exclusion of the Perez et al. & Marina et al. cohort resulted in a larger training sample size. These differences translated into substantial variation in model performance (RMSE), indicating that LODO estimates in this setting are strongly influenced by cohort-specific demographic structure rather than model robustness alone.

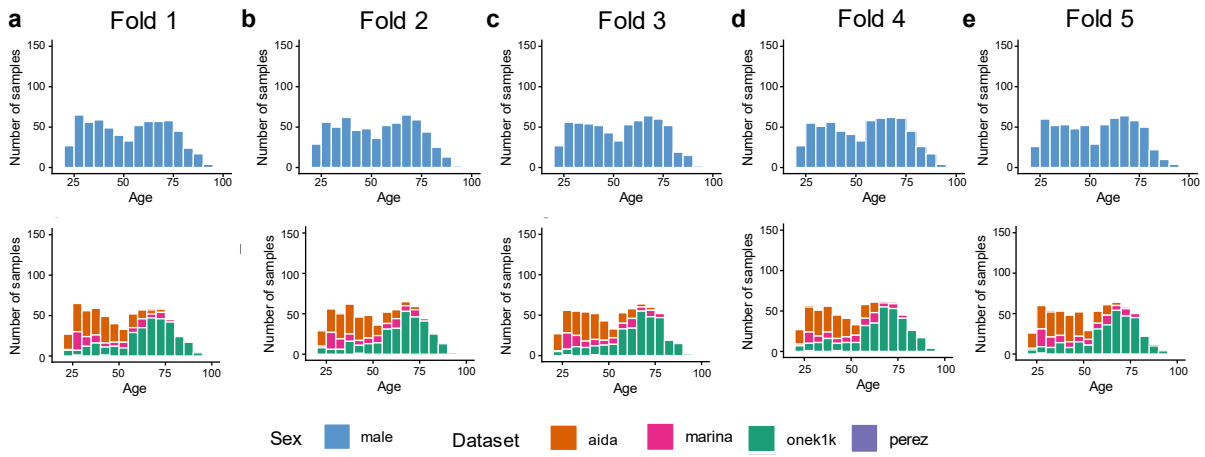


Figure R4. Age and dataset composition of five-fold cross-validation training sets for male MLP models. **a**-**e**, Age distribution and dataset composition of the five training folds used to build the male MLP models in the original analysis.

Fold	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Number of training samples	645	645	646	646	646
Mean age of training samples	53.27	53.49	53.71	54.23	53.61
Number of testing samples	162	162	161	161	161
Mean age of testing samples	55.23	54.35	53.48	51.38	53.86
Internal testing RMSE	8.59	9.39	8.91	8.28	8.23
External testing RMSE	16.26	15.71	14.94	16.24	15.62

Table R2. Sample distribution and model performance in five-fold cross-validation training sets for male MLP models.

For model development and comparison, we therefore employed stratified five-fold cross-validation, which preserves representation across age and sex groups in both training and testing folds. This approach provided substantially more balanced training and testing distributions and yielded more stable performance estimates across folds and across sex-stratified and combined models (Figure R4, Table R2, Figure R5).

Taken together, stratified five-fold cross-validation was used for model development and comparison, independent external cohorts were used to assess generalizability, and LODO was tested as an additional cohort-level sensitivity analysis.

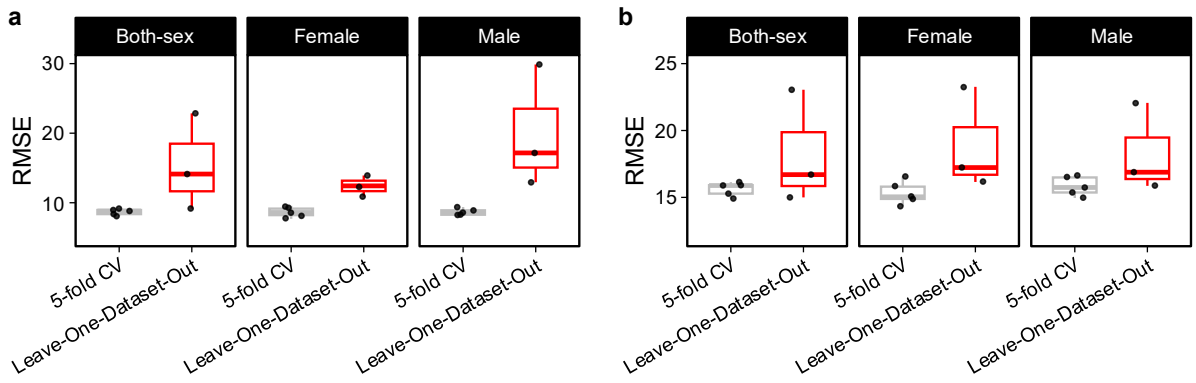


Figure R5. Comparison of model performance between five-fold cross-validation and leave-one-dataset-out (LODO) approaches. **a**-**b**, Root mean squared error (RMSE) of sex-combined and sex-stratified models trained using five-fold cross-validation or leave-one-dataset-out (LODO) and evaluated on the internal test set (**a**) and external test set (**b**).

Reviewer #1; Comment 9

9) Finally, aging clocks trained to predict chronological age have been increasingly criticized in the aging research community (e.g., see https://doi.org/10.1038/s43587-026-01066-6, https://doi.org/10.1038/s41467-025-66106-y). Second-generation clocks that predict clinically relevant outcomes such as mortality or multimorbidity are now generally preferred. While such outcome data may not be available in this study, the authors should explicitly acknowledge the limitations of first-generation clocks and discuss how this affects interpretation of their results.

Response

We thank the Reviewer for this important suggestion. According to the suggestion, we have now added a statement in the Discussion section (lines 494-498), explicitly acknowledging that our scRNA-seq-based aging clocks are first-generation models trained to predict chronological age, and therefore may not directly capture clinically relevant outcomes such as morbidity or mortality. We also highlight that integration with longitudinal clinical outcome data represents a promising direction for future studies.

Reviewer #2, General comments

In this study, Park et al. investigate the transcriptional changes in PBMCs in men and women during aging, using publicly available single cell RNA-seq from over 1800 individuals between 19 and 97 years old. The authors report distinct peaks of differential gene expression in different cell types (CD4 vs CD8) at different ages (40 vs 60), as well as sex-specific differences such as sex-dependent immune cell activation. They then apply machine learning based aging clocks to show that sex-stratified models outperform sex-combined models in learning from specific immune trajectories. The manuscript is well written and the findings are clearly described. While sex differences in immune aging are known and it is not surprising that aging clocks perform better on sex-stratified data, there is still a need for in depth

understanding of the sex differences and how they contribute to sex dimorphisms in disease. However, in the enclosed study, the robustness of the findings is difficult to assess due to lack of quality control information. In addition, the claims related to senescence are not supported. These points are described below.

Response

We would like to thank the Reviewer for his/her critical assessment of the paper. All of the Reviewer's points are valid and we take them into account in the revised version of the manuscript.

Reviewer #2; Comment 1

A major concern with the study is the uneven distribution of ages across the 4 public datasets. Most samples come from Onek1k and AIDA, which are enriched for an average age of ~40 (AIDA) and ~65 (Onek1k), which are the ages that appear to have the most DEFs. Further, since the individual datasets are clearly capturing either the first or second peak, it is plausible that the bimodal nature of the result is simply due to the uneven age distribution in the combined dataset.

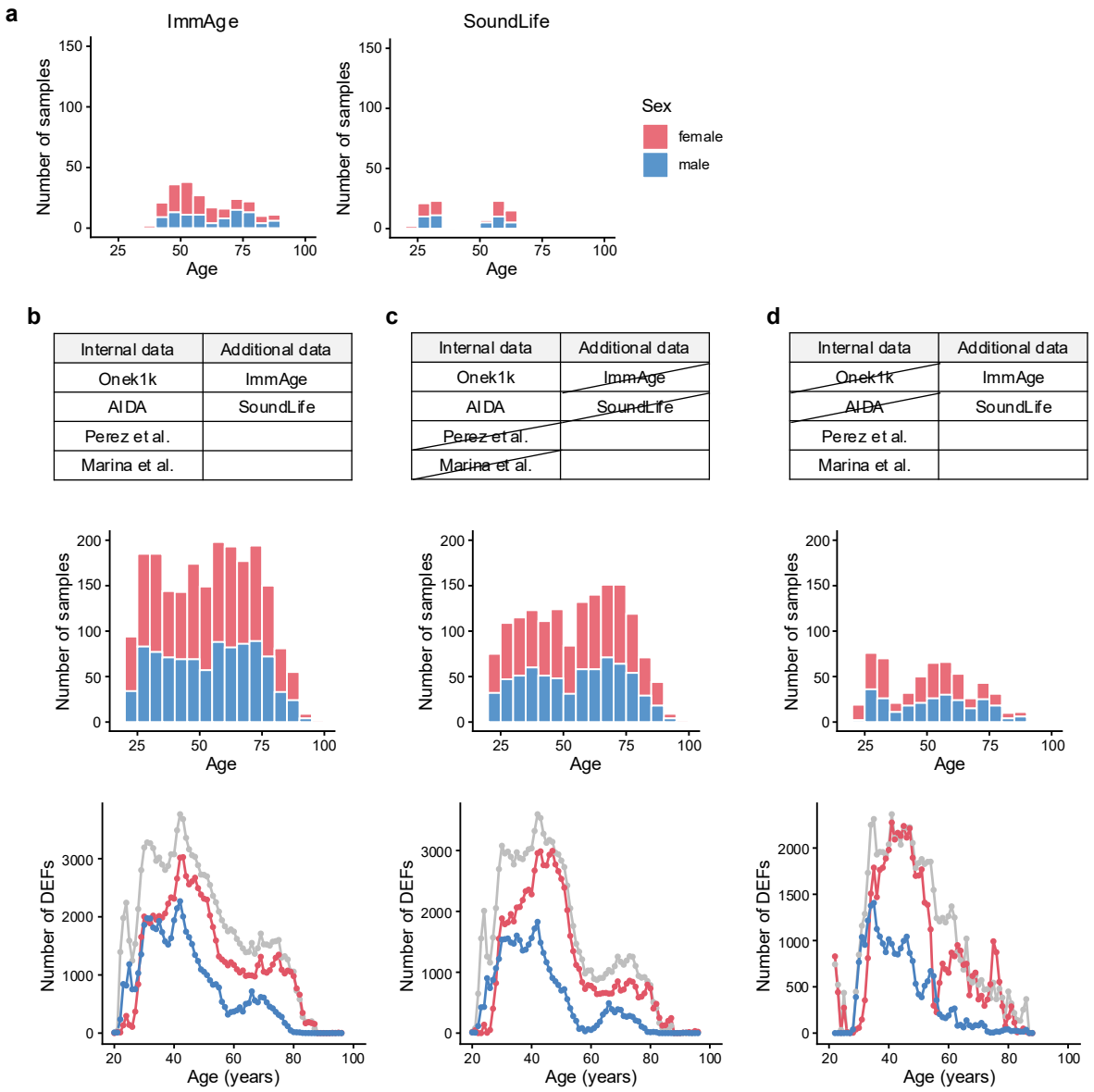
Response

This is a very important point and the Reviewer is correct by stating that the observed bimodal aging pattern could potentially arise from the uneven age distributions of the constituent datasets, particularly AIDA and OneK1K, which contribute a large proportion of samples and are enriched around approximately 40 and 65 years of age, respectively.

To evaluate whether this is the case, we opted for a step-by-step approach. We first expanded our analysis by incorporating two additional recently published PBMC datasets (ImmAge and SoundLife; see Methods lines 586-604; see Results lines 95-98). We then compared DE-SWAN results across three dataset combinations (Supplementary Figure 4 in the revised manuscript):

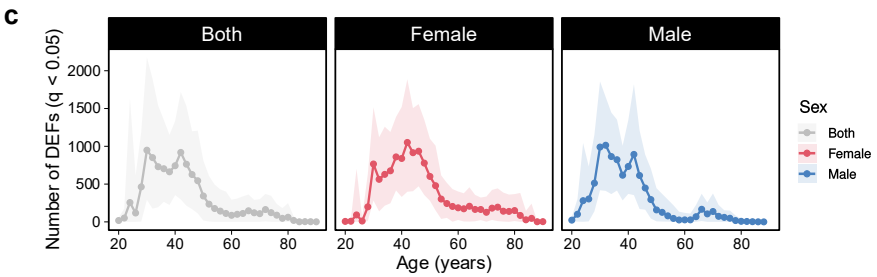
1. Original combined dataset plus the additional datasets
2. AIDA and OneK1K only
3. All datasets excluding AIDA and OneK1K, including the additional datasets.

The critical comparison that answers directly the Reviewer's point is between (2) and (3). As expected, the AIDA- and Onek1k-only analysis reproduced the prominent peaks around approximately 40 and 65 years of age. Crucially, when AIDA and Onek1k were completely excluded, the bimodal pattern remained evident, with peaks observed at similar age ranges. The second peak remained particularly pronounced in the female-specific analysis.



Supplementary Figure 4. Reproducibility of DE-SWAN peaks using additional PBMC scRNA-seq datasets. **a**, Age and sex distribution of the ImmAge and SoundLife datasets, which were used as additional PBMC scRNA-seq data for validation. **b-d**, Summary of included datasets, their age and sex distributions, and DE-SWAN results under different dataset combinations: all internal and additional datasets (**b**); Onek1k and AIDA datasets only (**c**); and Perez et al., Marina et al., and additional datasets (**d**).

Related to this concern, we also performed a modified DE-SWAN analysis with downsampling, as described in our response to Reviewer #1, Comment 6, to directly assess whether the observed peaks could be driven by differences in sample density across age. Specifically, for each age window tested (20-90 years, in 2-year increments), we randomly downsampled to 200 samples and repeated the analysis for 100 iterations. The bimodal pattern remained after downsampling, including the prominent first peak around 40 years of age and a smaller second peak around 65 years of age (most clearly observed in the male analysis; Extended Data Fig. 2c in the revised manuscript).



Extended Data Figure 2. DE-SWAN robustness analysis across parameters and cell types. **c**, Number of differentially expressed features (DEFs) across age identified by DE-SWAN using 100 random downsampling iterations, each using 200 samples.

Taken together, these complementary analyses demonstrate that the bimodal aging pattern is not solely driven by the age distributions of AIDA and Onek1k cohorts or by uneven sample density across age. The consistent observation of the two peaks across independent datasets and downsampling analyses supports the robustness and reproducibility of the identified aging transitions.

Reviewer #2; Comment 2

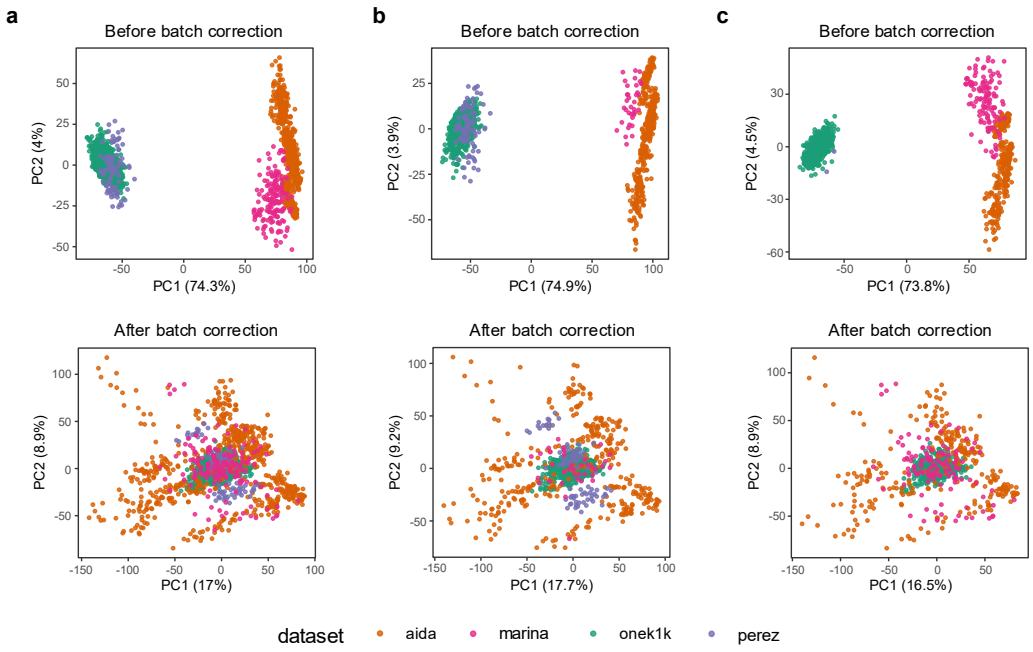
Related to the above point, although the authors include batch correction in the methods, there are no data (e.g. feature plots) showing batches before and after correction. This is important quality control, as it is possible that the different DE signatures in the two peaks are driven by differences in the AIDA and Onek1k datasets themselves.

Response

We thank the Reviewer for highlighting the importance of providing quality-control evidence for batch correction. This point overlaps with Reviewer #1's comments regarding dataset integration and batch effects (Reviewer #1, Comments #1-3).

As described in our response to Reviewer #1, dataset-associated variation was accounted for in the two downstream analyses. For DE-SWAN analysis, dataset identity was included as a covariate in the differential expression model, thereby directly controlling for dataset-specific effects during statistical testing. For clustering analysis, batch correction was performed on the merged pseudobulk expression matrices using limma's removeBatchEffect function prior to clustering and visualization.

To directly assess the effectiveness of batch correction, we have added PCA plots before and after correction in the revised manuscript (Supplementary Figure 4 in the revised manuscript). Prior to correction, samples exhibited partial separation by dataset, whereas this structure was substantially reduced after correction, indicating effective mitigation of dataset-associated variation.



Supplementary Figure 5. Batch effect correction for pseudobulk expression matrices. **a**-**c**, PCA plots before and after batch effect correction for both-sex (**a**), female (**b**), and male (**c**) pseudobulk expression matrices.

We additionally repeated the DE-SWAN analysis after applying removeBatchEffect to the pseudobulk expression matrix prior to differential expression analysis to evaluate whether the observed aging peaks were sensitive to the batch correction approach. The resulting age-associated patterns remained highly consistent with the original findings, including the presence and locations of the two major peaks and their associated cell-type signatures (Figure R1).

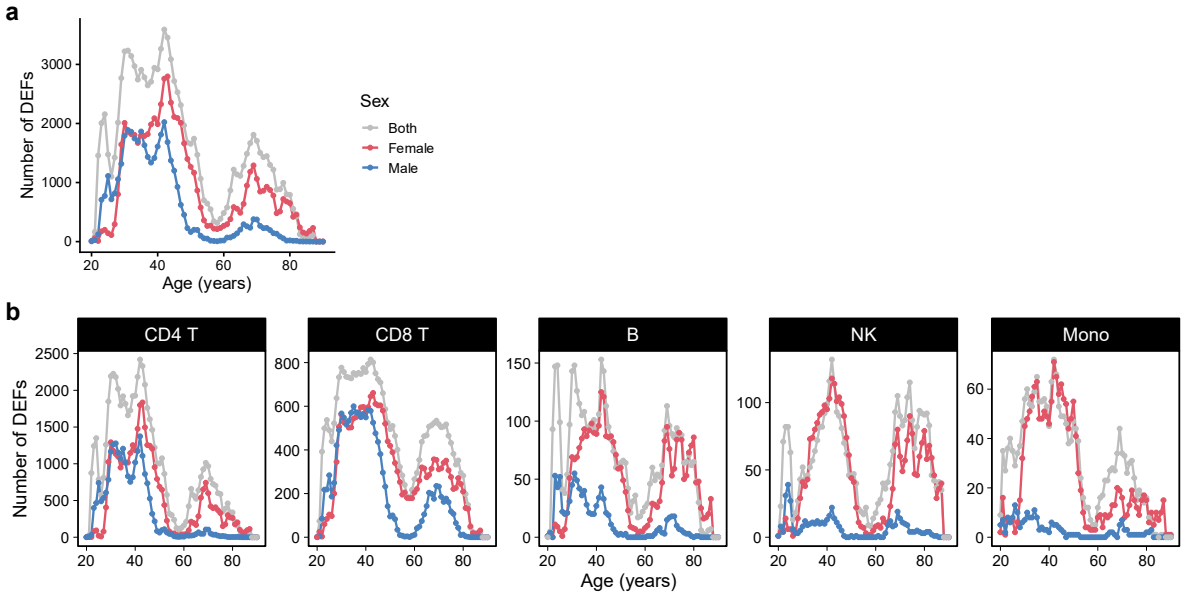


Figure R1. DE-SWAN results after applying removeBatchEffect. **a**-**b**, Number of differentially expressed features (DEFs, $q < 0.05$) across the age range identified by DE-SWAN algorithm after applying limma's removeBatchEffect, shown for all major PBMC cell types combined (**a**) and for each cell type separately (**b**).

Finally, the point that the two peaks may be driven by differences between the AIDA and OneK1K datasets is addressed above. Briefly, we showed that the bimodal pattern of aging remained evident even when AIDA and OneK1K were excluded from the analysis, with the two peaks observed at similar age ranges. Together, these analyses demonstrate that the observed bimodal aging pattern is not primarily driven by dataset-specific batch effects or the age distributions of the AIDA and OneK1K datasets.

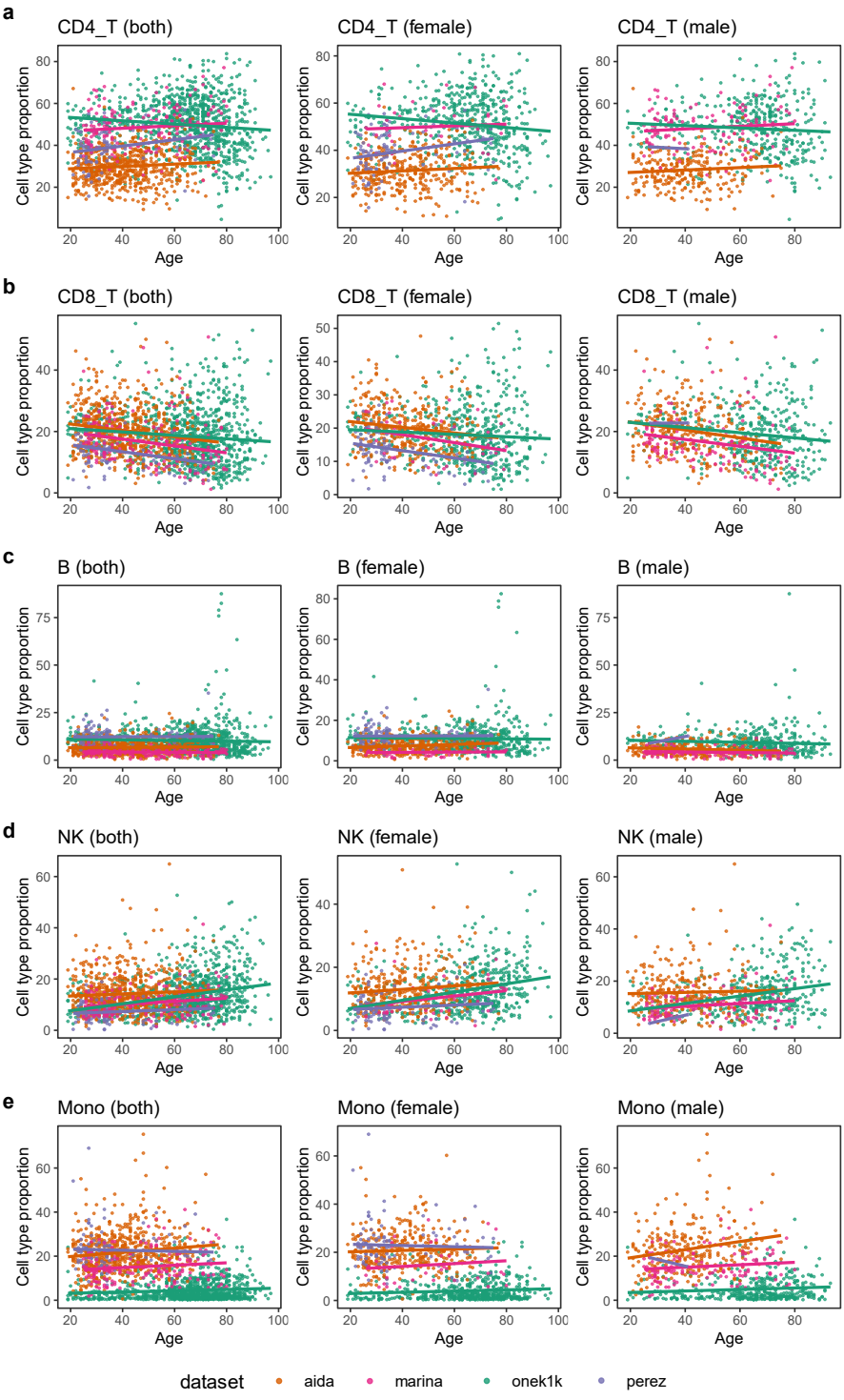
Reviewer #2; Comment 3

Data on the composition of each dataset should also be shown as this could bias the cell type analysis.

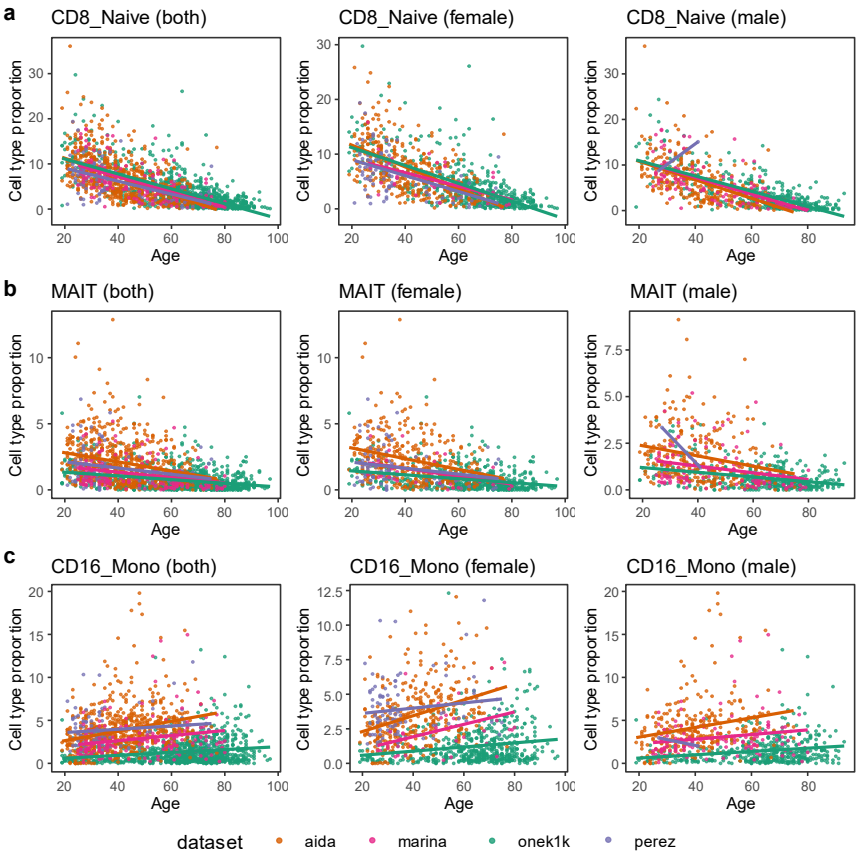
Response

We agree with the Reviewer. As described in our response to Reviewer #1 (Comment #4), we revised the cell type proportion analysis using the propeller framework (speckle R package), which is designed for differential cell type proportion analysis for scRNA-seq data, while accounting for biological replication and experimental covariates. In our analysis, dataset identity was included as a covariate to account for dataset-specific effects. The updated results are presented in the revised Table 1 (line 1255), and the relevant Results section has been revised accordingly (lines 72-75, lines 396-402).

To further assess the consistency of cell type proportion estimates across datasets, we generated dataset-stratified visualizations of cell type proportions. Supplementary Figures 2–3 in the revised manuscript show the proportions of the five major cell types (CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes) as well as the cell subtypes highlighted in the manuscript (lines 396-401; CD8 naïve T cells, MAIT cells, and CD16 monocytes). Similar age-associated trends were observed across the independent cohorts, supporting that the reported findings are not driven by a single dataset.



Supplementary Figure 2. Age-associated changes in major cell type proportions across the four datasets. **a**-**e**, Cell type proportions plotted against age for CD4 T cells (**a**), CD8 T cells (**b**), B cells (**c**), NK cells (**d**), and monocytes (**e**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.



Supplementary Figure 3: Age-associated changes in detailed cell subtype proportions across the four datasets. **a**-**c**, Cell type proportions plotted against age for CD8 Naïve T cells (**a**), MAIT cells (**b**), and CD16 monocytes (**c**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.

Reviewer #2; Comment 4

The claim that the female-specific late-increase cluster represents a senescence-associated cellular state is not supported. Senescent cells are rare, even in aged individuals and their gene expression signatures are well defined. There is no actual analysis of senescence in this study and this claim (included in the abstract) is purely speculative.

Response

The Reviewer is correct and we agree that we do not have direct evidence on a senescent-associated cell state. As such, we no longer refer to this state as senescent but with the observed pathways (i.e. apoptosis-related, iron accumulation). According to the Reviewer's suggestion, we have modified the statements referring to senescence in the abstract (line 25), introduction (line 60), results (lines 227-229, 244), and discussion sections (lines 446-450).

Reviewer #2; Comment 5

Minor:

Extended Fig. 4 is not readable.

Response

We thank the Reviewer for pointing this out. We have revised the figure and replaced it with a higher resolution version to improve clarity.

RESPONSE TO REVIEWERS' COMMENTS

First Round of Revision

Reviewer #1 (General comments)

The manuscript entitled “Sex-specific trajectories of nonlinear immune aging at single cell level” by Park et al. investigates sex-dependent dynamics of immune aging across the human lifespan using large-scale single-cell transcriptomic data. Using a multi-cohort integration of peripheral blood mononuclear cell (PBMC) scRNA-seq datasets and pseudobulk analytical frameworks, the authors report nonlinear trajectories of immune aging with distinct transcriptional inflection points around midlife (~40 years) and late life (>60 years), accompanied by sex-specific differences in particular functional pathways. The study is interesting and technically generally sound, but I have several major concerns that need to be addressed.

Response

We sincerely thank the Reviewer #1 for reading carefully the manuscript and providing highly constructive feedback. We have taken into account all of Reviewer #1’s suggestions and changed the manuscript which improved considerably.

Reviewer #1; Comments 1 to 3 (Batch effect correction)

1) Authors integrated and analyzed four large-scale scRNA-seq datasets (line 68) — they likely encountered dataset shift (batch effects). How exactly was this addressed? I would like to see a combined UMAP colored by dataset, which would demonstrate how well the datasets are mixed after integration. Currently, there is no clear description of dataset integration in the Methods. It almost appears that the datasets may have been analyzed separately — for example, pseudobulk values could have been calculated independently per dataset. This is not clearly stated. If this is the case, I am not convinced it is appropriate to state in the Results that “We integrated and analyzed four large-scale scRNA-seq datasets” (line 68), as the term “integrated” implies joint embedding or harmonization at the single-cell level. In the Methods, the authors instead state that “Pseudobulk expression matrices from all datasets were merged” (line 721). It is unclear what “merged” means in this context — were values averaged, concatenated, or simply combined across individuals? A more precise description is required. Moreover, the issue of batch effects remains: even at the pseudobulk level, the authors should demonstrate that dataset-specific biases are adequately controlled.

2) The authors mention that “The merged pseudobulk expression matrix was corrected for dataset-specific batch effects using removeBatchEffect function from limma” (line 740), but this statement appears after the “Sex-stratified differential expression sliding window analysis (DE-SWAN)” section (line 723) in the Methods. Does this imply that DE-SWAN was performed on uncorrected data? If so, this would be a significant methodological flaw, as batch effects could strongly bias differential expression results.

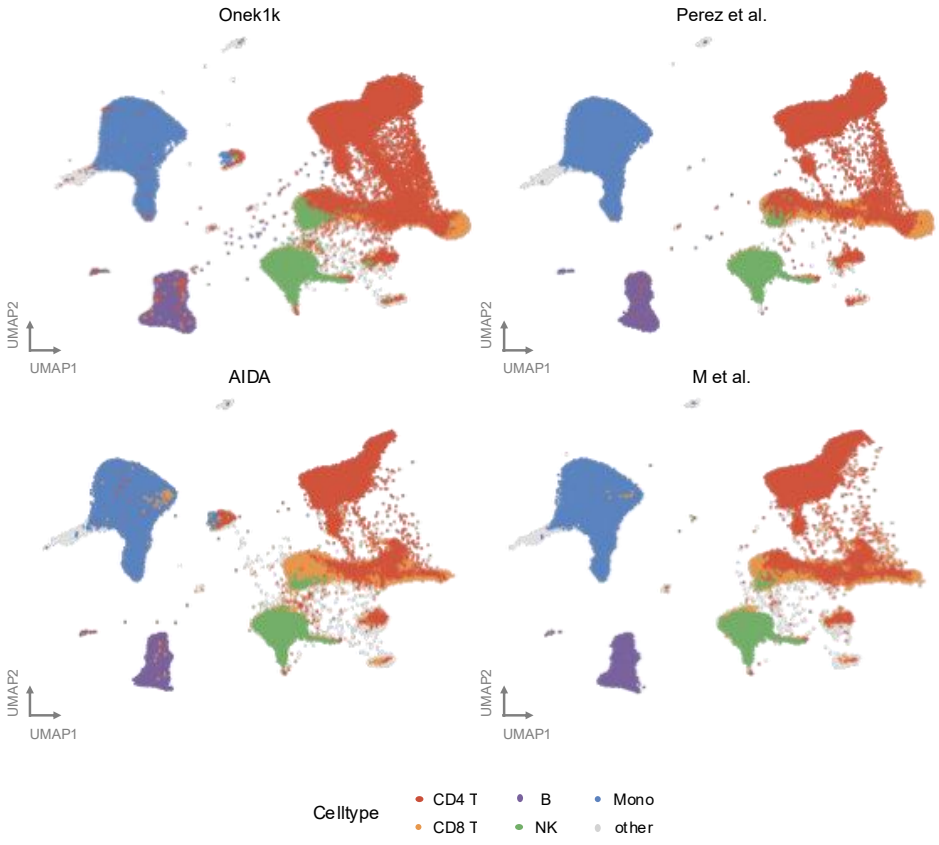
3) In any case, I would like to see diagnostic evidence of batch effect correction at the pseudobulk level — for example, a PCA plot of samples colored by dataset before and after correction, demonstrating that samples do not cluster by dataset after correction. This is standard practice and necessary for validating batch effect correction quality.

Response

We thank the Reviewer for raising these important points regarding dataset integration and batch effect correction. We agree with the Reviewer that these steps were insufficiently described in the original manuscript and have now clarified both the analytical workflow (lines 737-738, 750, 758-759, 780-783) and the terminology used throughout the manuscript (lines 19, 51, 55-56, 79, 87, 110, 387, 490, 504, 511, 1232-1235, 1341-1342).

The Reviewer is correct that the term “integrated” incorrectly implies joint embedding or harmonization at the single-cell level and we have revised the manuscript accordingly to avoid this interpretation. To further clarify, our study did not perform single-cell-level integration or harmonization using methods such as Seurat integration, Harmony, or scVI. Instead, each dataset was independently projected onto the Azimuth human PBMC reference to obtain consistent cell type annotations across datasets. Based on these annotations, donor-level pseudobulk expression matrices were generated separately for each dataset and subsequently combined by concatenating matched gene features across donors for downstream analyses.

In line with the Reviewer’s suggestion, we demonstrate the consistency of cell type annotation across all datasets through a UMAP plot colored by major immune cell types and split by dataset (Supplementary Figure 1 in the revised manuscript). This visualization shows that corresponding immune cell populations from the four datasets map consistently onto the PBMC reference space.



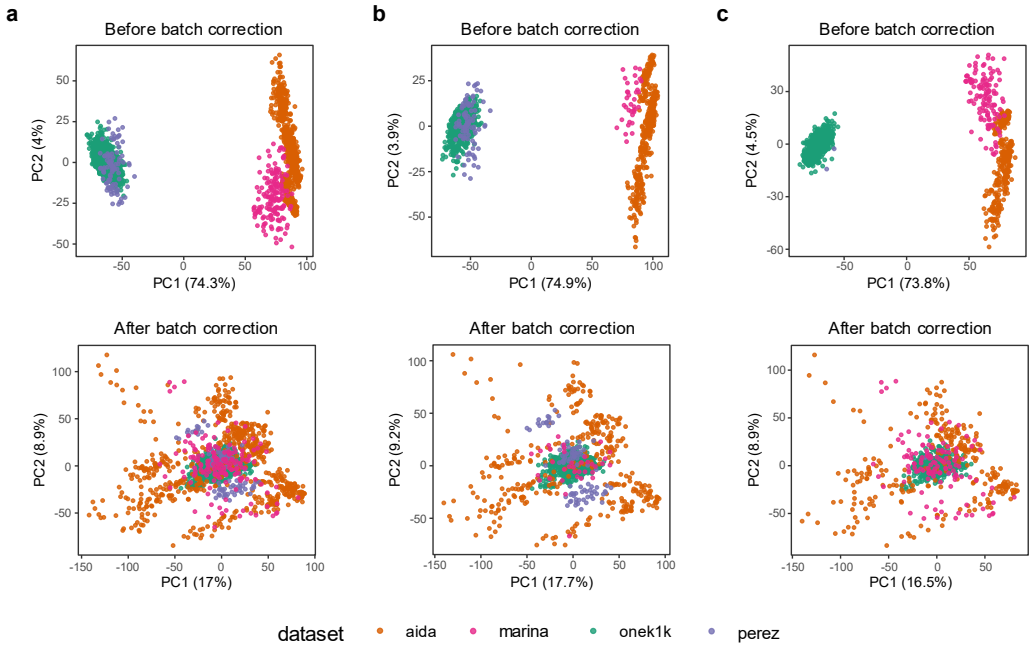
Supplementary Figure 1. UMAP projection of major PBMC cell types across the four datasets.

Regarding batch effects, we accounted for dataset-specific variation separately in the two main downstream analyses.

Importantly, we would like to clarify that DE-SWAN was not performed on uncorrected data. Dataset identity was included as a covariate in the linear modeling framework used for differential expression testing. Batch-associated variation was therefore controlled directly within the regression model during differential expression analysis. According to the Reviewer's suggestion, we have revised the Methods section to clarify this point more explicitly (lines 758-759).

Second, for the downstream clustering analysis of significant age-associated features (SAFs) identified by DE-SWAN, we applied limma's removeBatchEffect function to the merged pseudobulk expression matrices (performed separately for both-sex, female, and male analyses) prior to trajectory clustering. In this context, batch correction was applied specifically for clustering and visualization purposes.

As requested by the Reviewer, we now include PCA plots before and after batch effect correction at the pseudobulk level (Supplementary Figure 5 in the revised manuscript). Prior to correction, samples exhibited partial separation by dataset, whereas this dataset-associated structure was substantially reduced after correction, indicating effective mitigation of batch effects.



Supplementary Figure 5. Batch effect correction for pseudobulk expression matrices. **a**-**c**, PCA plots before and after batch effect correction for both-sex (**a**), female (**b**), and male (**c**) pseudobulk expression matrices.

We additionally evaluated whether applying removeBatchEffect prior to DE-SWAN altered the results. The overall findings remained highly consistent with the original analysis (Figure R1), including the presence of two major peaks in both sexes, larger peak magnitudes in females, enrichment of CD4 T cell features in the first peak around 40 years of age, and increased contribution of CD8 T cell features in the second peak after approximately 60 years of age.

These results demonstrate that primary conclusions are robust to alternative batch correction strategies.

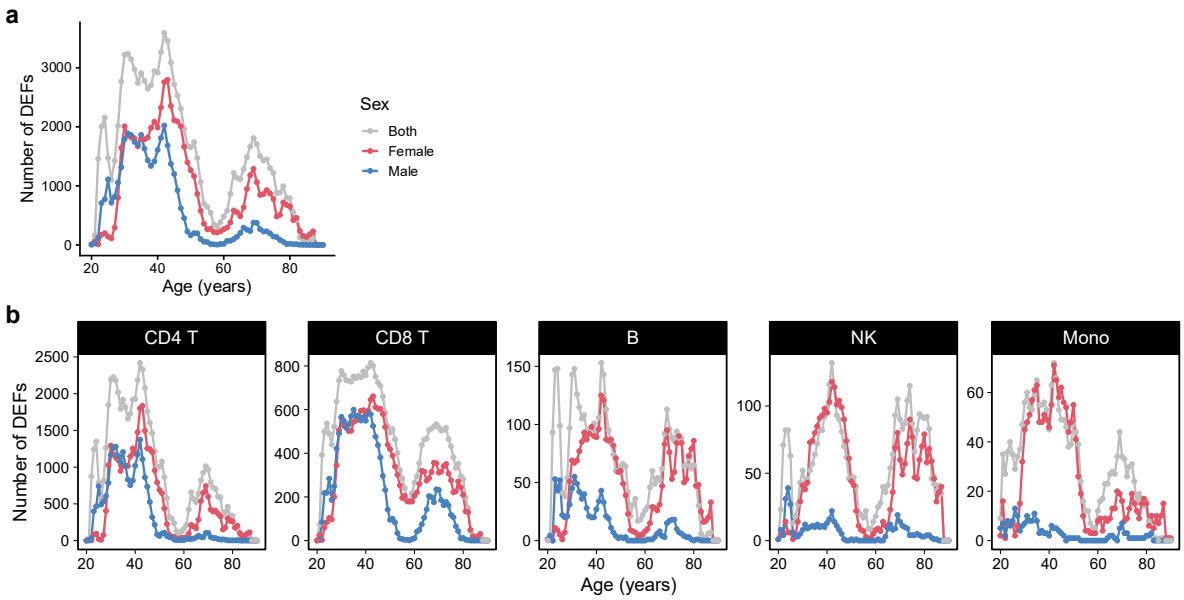


Figure R1. DE-SWAN results after applying removeBatchEffect. **a**-**b**, Number of differentially expressed features (DEFs, $q < 0.05$) across the age range identified by DE-SWAN algorithm after applying limma's removeBatchEffect, shown for all major PBMC cell types combined (**a**) and for each cell type separately (**b**).

Reviewer #1; Comment 4

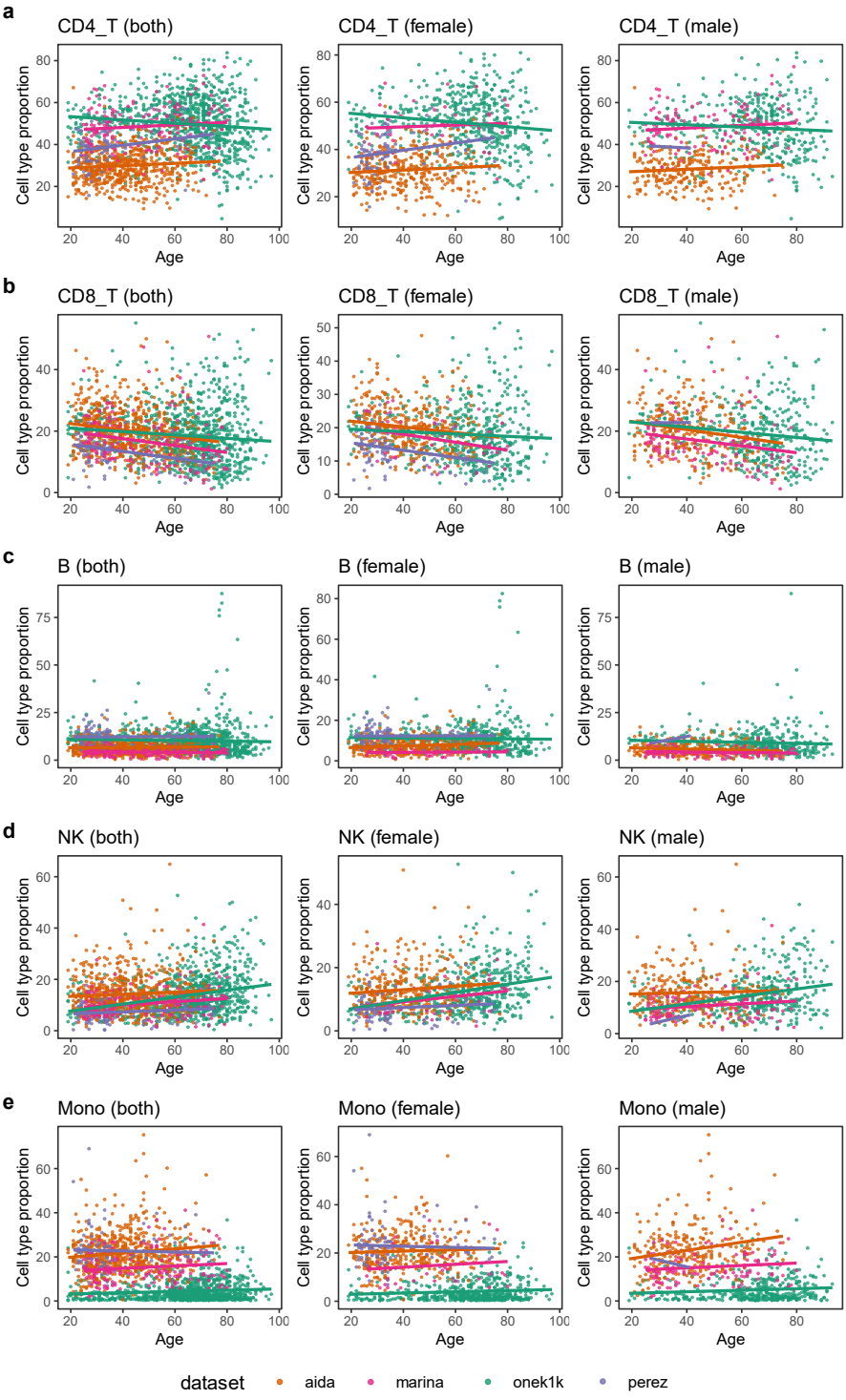
4) The authors performed cell type proportion analyses (line 72) based on scRNA-seq data — however, such estimates are known to be biased due to technical artifacts including differential capture efficiency and cell-type-specific dissociation sensitivity. These issues are well recognized in the field, and dedicated computational frameworks have been developed to mitigate them (e.g., see https://doi.org/10.1093/bioinformatics/btac582). Could the authors validate their findings using an orthogonal method (e.g., flow cytometry or deconvolution of bulk RNA-seq), or at least discuss these limitations? At a minimum, since multiple independent datasets are available, the authors should treat them as biological replicates and demonstrate that inferred cell type proportions are consistent across datasets.

Response

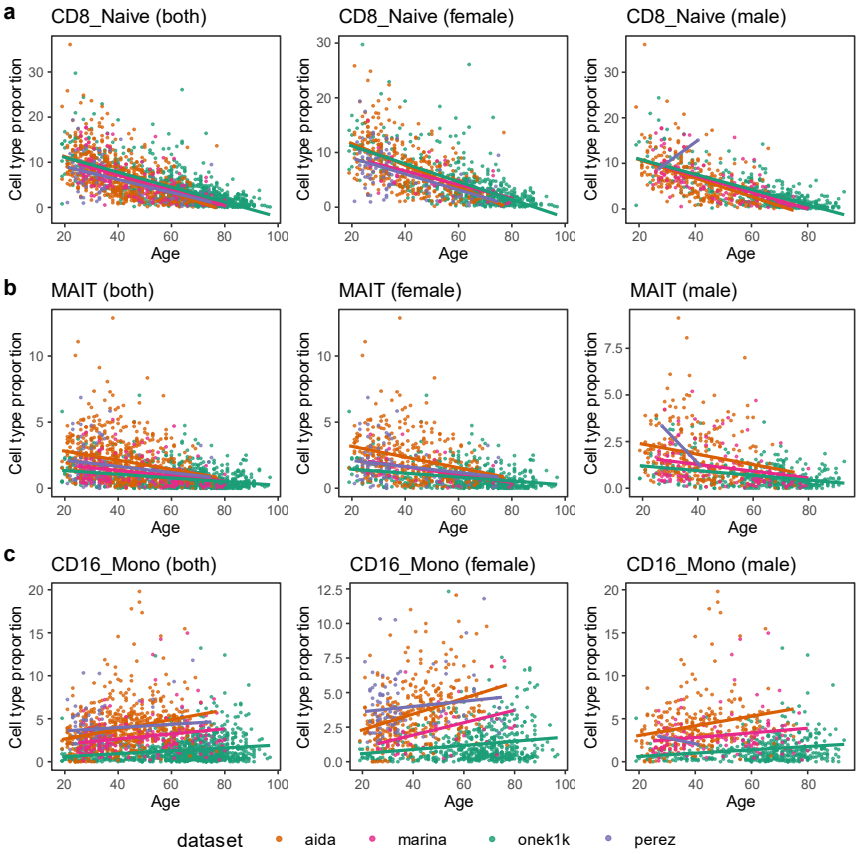
We agree with the Reviewer's assessment on the potential technical biases in scRNA-seq-based cell type proportion analyses. The Reviewer is correct that inferred cell type proportions can be influenced by dataset-specific factors, including differences in cell capture efficiency, dissociation sensitivity, and sample processing.

As suggested by the Reviewer, we revised the cell type proportion analysis using the propeller framework (speckle R package). Propeller is specifically designed for differential cell type proportion analysis while accounting for biological replication and experimental covariates. Instead of assessing simple correlations between age and cell type proportions in the merged dataset, we modelled donor-level cell type proportions using a linear framework with dataset identity included as a covariate. The updated results are presented in the revised Table 1 (line 1472), and the relevant Results and Methods section has been revised accordingly (lines 83-86, lines 511-514, lines 833-840).

In addition, to evaluate the consistency of cell type proportion estimates across datasets, we generated dataset-stratified visualizations of cell type proportions across age. Supplementary Figures 2–3 in the revised manuscript show the proportions of the five major cell types (CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes), as well as the cell subtypes highlighted in the manuscript (lines 511-514; CD8 naïve T cells, MAIT cells, and CD16 monocytes). Overall, similar age-associated trends were observed across the independent cohorts, supporting that the findings are not driven by a single dataset.



Supplementary Figure 2. Age-associated changes in major cell type proportions across the four datasets. **a**-**e**, Cell type proportions plotted against age for CD4 T cells (**a**), CD8 T cells (**b**), B cells (**c**), NK cells (**d**), and monocytes (**e**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.



Supplementary Figure 3: Age-associated changes in detailed cell subtype proportions across the four datasets. **a**-**c**, Cell type proportions plotted against age for CD8 Naïve T cells (**a**), MAIT cells (**b**), and CD16 monocytes (**c**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.

To follow up on the Reviewer's suggestion, we have now explicitly discussed the limitations of scRNA-seq-based cell type proportion inference and the potential influence of technical variability in the Discussion section (lines 621-624).

Reviewer #1; Comment 5

5) The authors generated pseudobulk expression profiles for five major PBMC cell types: CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes (line 79). However, proportions of subtypes within these major populations (e.g., naive vs memory T cells) are known to change substantially with age and could confound the observed transcriptional changes. Could the authors demonstrate that subtype composition remains relatively stable across age, or alternatively perform analyses at a finer cell subtype resolution?

Response

This is indeed a crucial point. We agree with the Reviewer that substantial shifts in immune subtypes occur with aging, particularly within T cell compartments, where naïve CD4 and CD8 T cells decrease while memory and effector populations increase. Consistent with this, our updated cell type proportion analysis (Table 1) also shows significant age-associated changes in these subpopulations. This raises the possibility that transcriptional changes observed in aggregated CD4 and CD8 T cell populations could be driven by shifts in subtype composition.

To address the Reviewer's point, we performed DE-SWAN analysis at a finer cellular resolution for CD4 and CD8 T cells that exhibited strong signals in the original analysis (Figure R2). Specifically, we analyzed CD4 naïve T cells (CD4 Naïve), CD4 central memory T cells (CD4 TCM), CD4 effector memory T cells (CD4 TEM), CD8 naïve T cells (CD8 Naive), CD8 central memory T cells (CD8 TCM), and CD8 effector memory T cells (CD8 TEM).

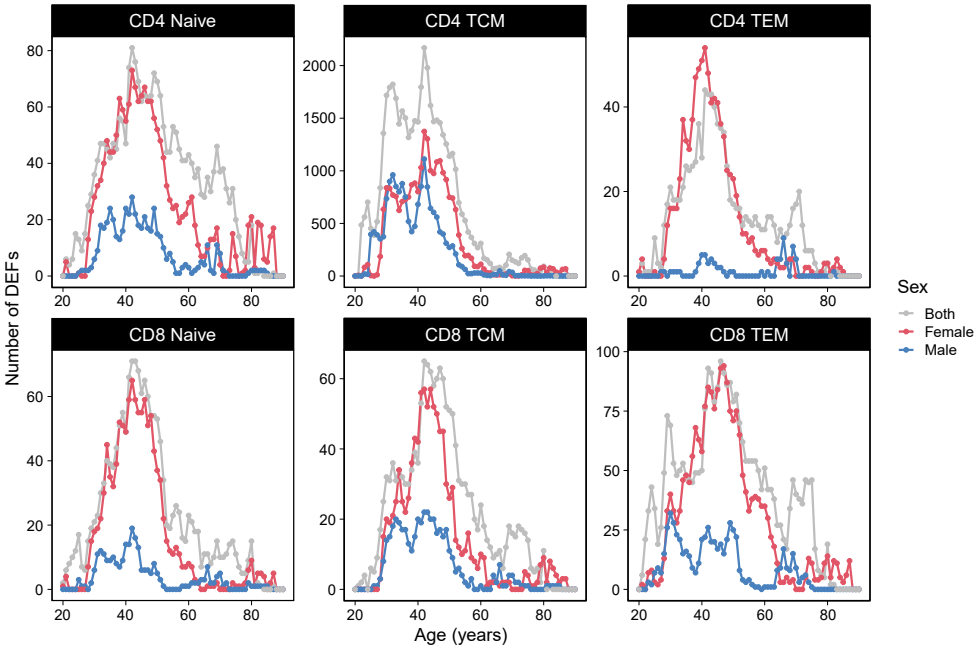


Figure R2. DE-SWAN results of CD4 and CD8 T cell subtypes. Number of differentially expressed features (DEFs, $q < 0.05$) across the age range identified by DE-SWAN algorithm in CD4 naïve T cells (CD4 Naïve), CD4 central memory T cells (CD4 TCM), CD4 effector memory T cells (CD4 TEM), CD8 naïve T cells (CD8 Naive), CD8 central memory T cells (CD8 TCM), and CD8 effector memory T cells (CD8 TEM).

Notably, the results show that the major transcriptional aging patterns observed in aggregated CD4 and CD8 T-cell populations are largely preserved at the subtype level. In particular, a prominent peak of differentially expressed features around ~40 years of age is consistently observed across multiple T-cell subtypes, with a second, weaker peak after ~60 years also detectable in several subtypes, particularly in the both-sex analysis (Figure R2).

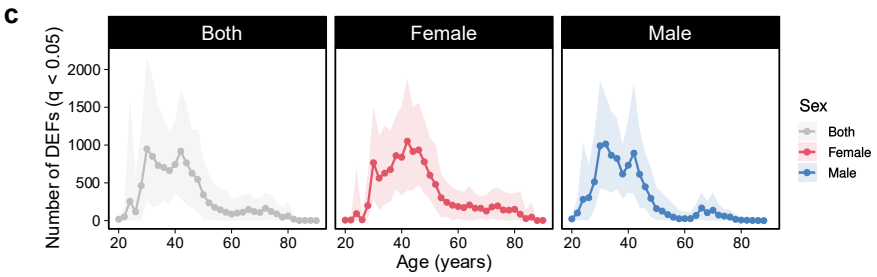
Together, these results indicate that the transcriptional aging signatures identified in aggregated CD4 and CD8 T-cell populations are not solely driven by shifts in subtype composition. Rather, similar nonlinear aging patterns are detectable within individual T-cell subtypes, supporting the robustness of the observed age-associated transcriptional trajectories.

Reviewer #1; Comment 6

6) The authors report two peaks of differential expression — around ~40 years and after 60 years (line 89). These peaks appear to coincide with higher sample numbers in these age ranges. Could this observation be partially driven by increased statistical power rather than true biological effects? The authors should control for sample size effects, for example by downsampling or applying methods that account for uneven age distribution.

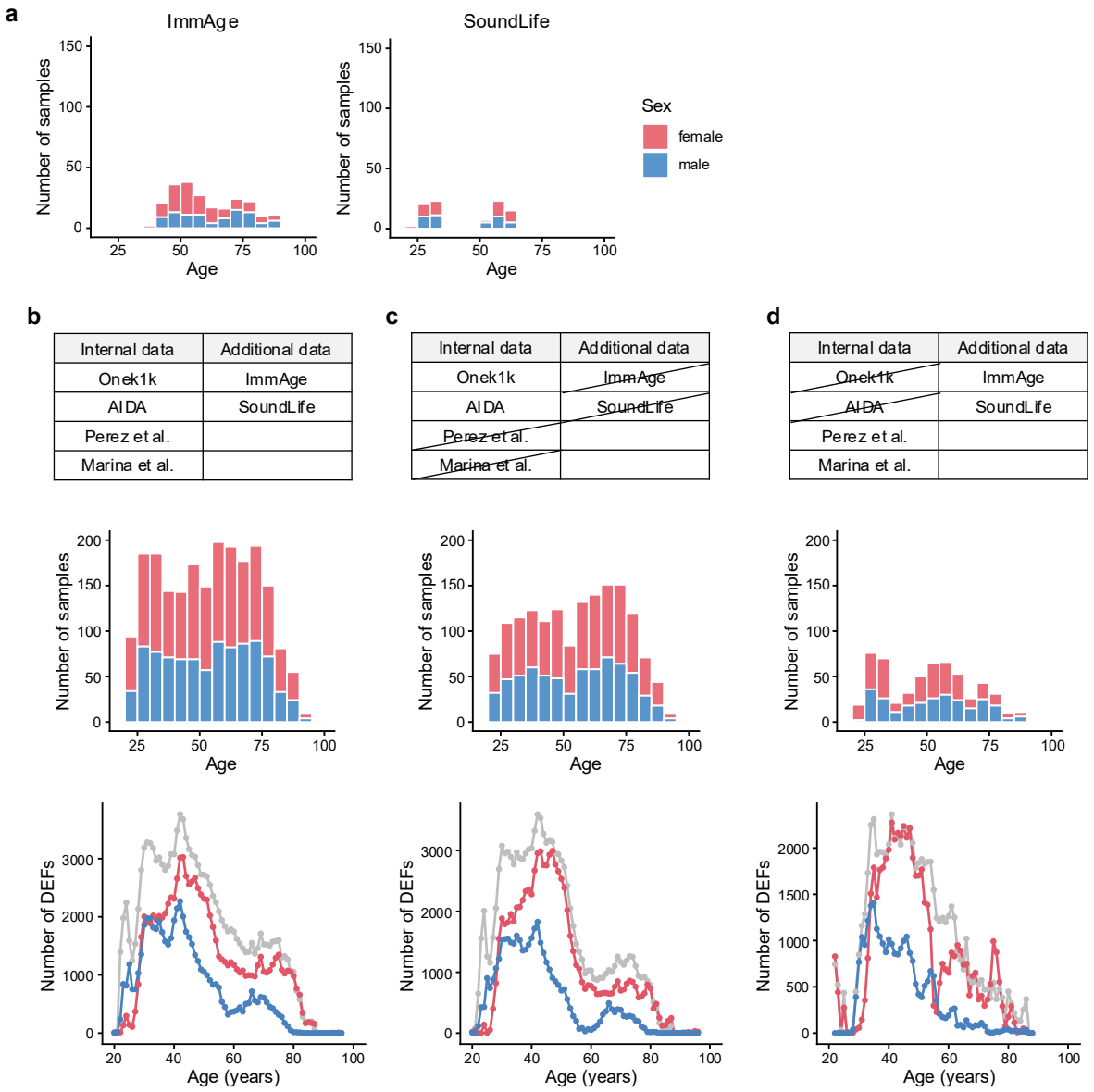
Response

The Reviewer is correct to suggest that the peaks of differential expression may be driven by sample distribution. To address the Reviewer's concern, we performed a modified DE-SWAN analysis with downsampling as suggested by the Reviewer and added the results as Extended Data Figure 2c in the revised manuscript. Specifically, for each age window tested (20-90 years, in 2-year increments), we randomly downsampled to 200 samples and repeated the analysis for 100 iterations. The two peaks remained after downsampling, including the prominent first peak around 40 years of age and a smaller second peak around 65 years of age (most clearly observed in the male analysis). We have added the corresponding methods and results descriptions to the manuscript (lines 767-777, 102-103).



Extended Data Figure 2. DE-SWAN robustness analysis across parameters and cell types. **c**, Number of differentially expressed features (DEFs) across age identified by DE-SWAN using 100 random downsampling iterations, each using 200 samples.

We further evaluated whether the observed bimodal pattern could be driven by the age distributions of the major contributing datasets, AIDA and OneK1K, as discussed in our response to Reviewer #2, Comment #1. Briefly, we incorporated two additional recently published PBMC datasets (ImmAge and SoundLife) and repeated the DE-SWAN analysis using different combinations of datasets (Supplementary Figure 4 in the revised manuscript). Importantly, the bimodal pattern remained evident even when AIDA and OneK1K were completely excluded from the analysis, with peaks observed at similar age ranges. Together, the downsampling analysis and the reproducibility of the two peaks across independent dataset combinations support the conclusion that these aging transitions are unlikely to be explained solely by sample size differences or uneven age distributions and instead reflect underlying biological aging processes.



Supplementary Figure 4. Reproducibility of DE-SWAN peaks using additional PBMC scRNA-seq datasets. **a**, Age and sex distribution of the ImmAge and SoundLife datasets, which were used as additional PBMC scRNA-seq data for validation. **b-d**, Summary of included datasets, their age and sex distributions, and DE-SWAN results under different dataset combinations: all internal and additional datasets (**b**); Onek1k and AIDA datasets only (**c**); and Perez et al., Marina et al., and additional datasets (**d**).

Reviewer #1; Comment 7

7) Which gene background was used for functional enrichment analyses? No information is provided in the Methods (lines 769–778), suggesting that the default background may have been used. This is problematic: enrichment analyses should use expressed genes within each cell type as background. Otherwise, results may largely reflect general cell-type-specific functions rather than trajectory-specific biological processes.

Response

In the original analysis, we used all genes represented in the pseudobulk expression matrix as the background set. This matrix was constructed using features (cell type-gene pairs; cell type-specific expression) that were expressed across all donors in all four datasets (>0 pseudobulk expression in all donors) as mentioned in the Methods section (lines 751-752). While this ensures that only consistently detected genes were included, we agree with the Reviewer that using a cell type-agnostic background is suboptimal because it does not account for cell-type-specific gene expression.

To address this point, we have revised the pathway enrichment analysis using cell-type-specific background gene sets, defined as genes (features) expressed in each cell type across all donors. This approach better accounts for differences in baseline gene expression between cell types and reduces bias toward general cell-type-specific functions.

We have updated the Methods part accordingly (lines 811-813) and revised all affected figures (Fig. 3h, Fig. 3j, Fig. 3m, Extended Data Fig. 4, Extended Data Fig. 5a, Extended Data Fig. 5c, Extended Data Fig. 5g) and results sections (lines 182-186, 209-280) to reflect the corrected analysis. Importantly, the revised analysis yielded consistent biological interpretations, with no change to the main conclusions of the functional enrichment results.

Reviewer #1; Comment 8

8) Given that four independent datasets were used, it would be more appropriate to train and test the aging clock models across datasets (e.g., leave-one-dataset-out validation) rather than relying solely on five-fold cross-validation (line 817). This would provide a more robust assessment of generalizability and reduce the risk of overfitting.

Response

We agree that leave-one-dataset-out (LODO) validation can provide a stringent assessment of generalizability across cohorts. However, we would like to clarify that, in the present study, the datasets differ substantially in sample size and age-sex composition. As such, LODO validaton in this setting introduces confounding factors such as demographic structure, making performance dependent on the excluded cohort, rather than reflecting intrinsic model generalizability. This consideration was important for our analysis, as we aimed to (i) compare and select appropriate algorithms (elastic net regression, XGBoost, and MLP) and (ii) evaluate performance differences between sex-combined and sex-stratified models.

To illustrate this, we performed LODO validation using MLP models. Because of strong sex imbalance between Perez et al. (97.9% female) and Marina et al. (77.6% male) (Extended Data Fig. 1b), these two datasets were combined into a single cohort for this analysis, resulting in three effective datasets: Onek1k, AIDA, and the combined Perez/Marina cohort.

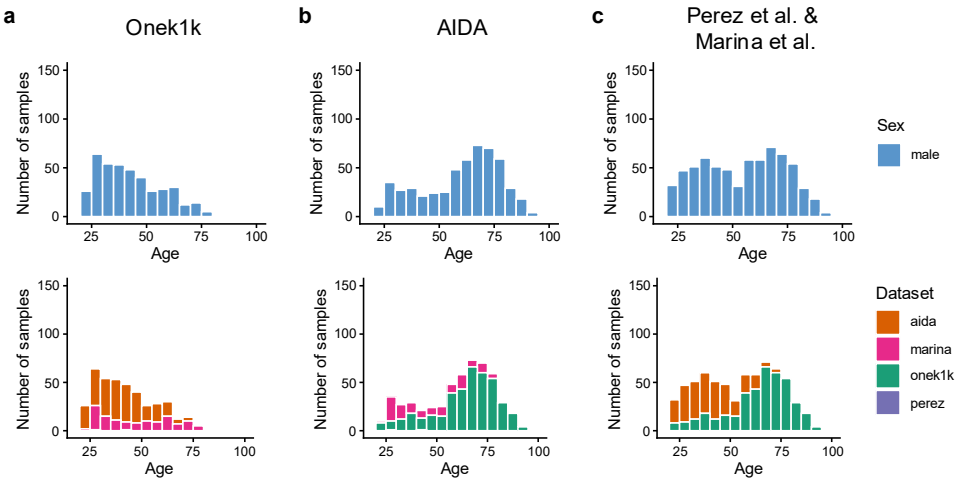


Figure R3. Age and dataset composition of leave-one-dataset-out (LODO) training sets for male MLP models. **a-c**, Age distribution and dataset composition of the LODO training sets when the Onek1k dataset was held out (**a**), the AIDA dataset was held-out (**b**), and the Perez et al. and Marina et al. datasets were held out (**c**).

Held-out dataset	Onek1k	AIDA	Perez et al. & Marina et al.
Number of training samples	879	1206	1571
Mean age of training samples	43.06	60.18	54.81
Number of testing samples	949	622	257
Mean age of testing samples	64.14	40.98	47.54
Internal testing RMSE	29.88	17.16	12.95
External testing RMSE	22.06	16.88	15.86

Table R1. Sample distribution and model performance in leave-one-dataset-out (LODO) training sets for male MLP models.

As shown in the LODO results, the composition of the training set varied substantially depending on the held-out dataset (Figure R3, Table R1). For example, when Onek1k was excluded, the training set contained a smaller sample size with a narrower and younger age distribution, whereas exclusion of AIDA cohort resulted in a markedly older training distribution. In contrast, exclusion of the Perez et al. & Marina et al. cohort resulted in a larger training sample size. These differences translated into substantial variation in model performance (RMSE), indicating that LODO estimates in this setting are strongly influenced by cohort-specific demographic structure rather than model robustness alone.

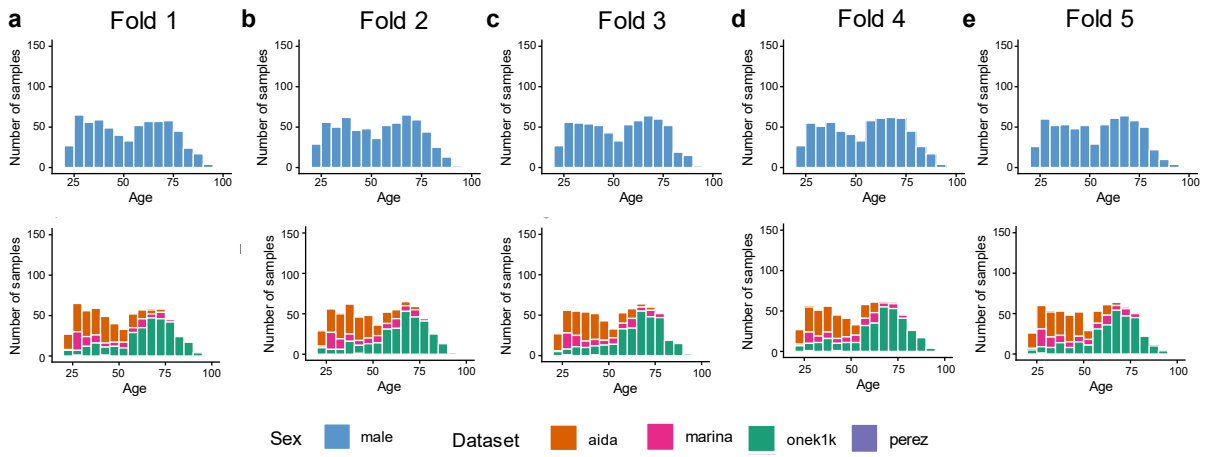


Figure R4. Age and dataset composition of five-fold cross-validation training sets for male MLP models. **a**-**e**, Age distribution and dataset composition of the five training folds used to build the male MLP models in the original analysis.

Fold	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Number of training samples	645	645	646	646	646
Mean age of training samples	53.27	53.49	53.71	54.23	53.61
Number of testing samples	162	162	161	161	161
Mean age of testing samples	55.23	54.35	53.48	51.38	53.86
Internal testing RMSE	8.59	9.39	8.91	8.28	8.23
External testing RMSE	16.26	15.71	14.94	16.24	15.62

Table R2. Sample distribution and model performance in five-fold cross-validation training sets for male MLP models.

To mitigate the above-mentioned challenges, we employed five-fold cross-validation within the training data to obtain a more stable and balanced estimate of model performance. This approach preserves representation across age and sex strata in both training and testing folds, reducing the risk that performance is driven by a single cohort with skewed composition. As shown in the cross-validation results, training and testing folds we used in the original analysis exhibited consistent age and sex distributions and more stable performance estimates across folds (Figure R4, Table R2).

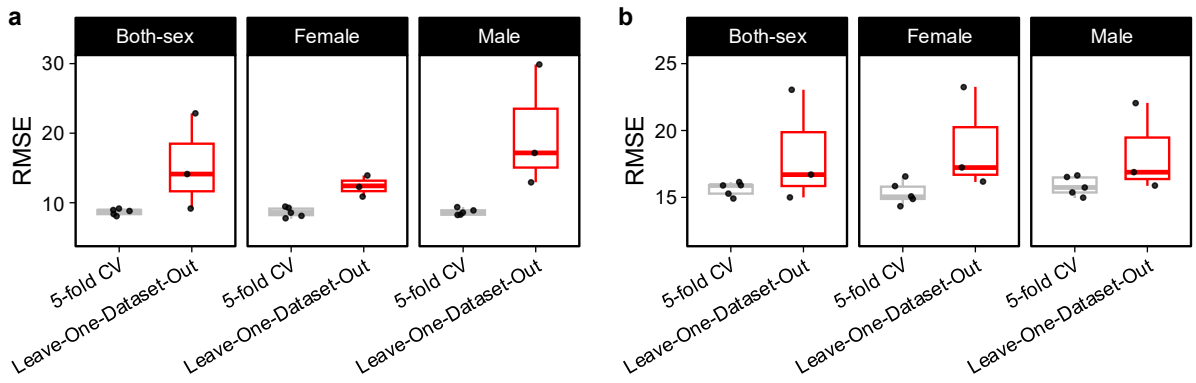


Figure R5. Comparison of model performance between five-fold cross-validation and leave-one-dataset-out (LODO) approaches. **a**-**b**, Root mean squared error (RMSE) of sex-combined and sex-stratified models trained using five-fold cross-validation or leave-one-dataset-out (LODO) and evaluated on the internal test set (**a**) and external test set (**b**).

Importantly, while LODO showed high variability due to cohort composition, the five-fold cross-validation results were consistent across folds and across sex-stratified and combined models (Figure R5), supporting the stability of model training under balanced sampling.

Finally, we agree with the reviewer that generalizability beyond the discovery datasets is an important consideration in aging clock models. We therefore emphasize that the final models were evaluated on two additional independent datasets in the revised manuscript (ImmAge and SoundLife; lines 717-735, line 450), which were not used in model training or parameter tuning. In the ImmAge cohort, correlations were 0.69 (both-sex), 0.78 (female), and 0.80 (male). In the SoundLife cohort, correlations were 0.80 (both-sex), 0.78 (female), and 0.65 (male). These results reinforce the robustness and generalizability of the proposed models across additional unseen populations and further strengthen the external validation results originally presented in the submitted manuscript.

Reviewer #1; Comment 9

9) Finally, aging clocks trained to predict chronological age have been increasingly criticized in the aging research community (e.g., see https://doi.org/10.1038/s43587-026-01066-6, https://doi.org/10.1038/s41467-025-66106-y). Second-generation clocks that predict clinically relevant outcomes such as mortality or multimorbidity are now generally preferred. While such outcome data may not be available in this study, the authors should explicitly acknowledge the limitations of first-generation clocks and discuss how this affects interpretation of their results.

Response

We thank the Reviewer for this important suggestion. According to the suggestion, we have now added a statement in the Discussion section (lines 625-629), explicitly acknowledging that our scRNA-seq-based aging clocks are first-generation models trained to predict chronological age, and therefore may not directly capture clinically relevant outcomes such as morbidity or mortality. We also highlight that integration with longitudinal clinical outcome data represents a promising direction for future studies.

Reviewer #2, General comments

In this study, Park et al. investigate the transcriptional changes in PBMCs in men and women during aging, using publicly available single cell RNA-seq from over 1800 individuals between 19 and 97 years old. The authors report distinct peaks of differential gene expression in different cell types (CD4 vs CD8) at different ages (40 vs 60), as well as sex-specific differences such as sex-dependent immune cell activation. They then apply machine learning based aging clocks to show that sex-stratified models outperform sex-combined models in learning from specific immune trajectories. The manuscript is well written and the findings are clearly described. While sex differences in immune aging are known and it is not surprising that aging clocks perform better on sex-stratified data, there is still a need for in depth understanding of the sex differences and how they contribute to sex dimorphisms in disease. However, in the enclosed study, the robustness of the findings is difficult to assess due to lack of quality control information. In addition, the claims related to senescence are not supported. These points are described below.

Response

We would like to thank the Reviewer for his/her critical assessment of the paper. All of the Reviewer's points are valid and we take them into account in the revised version of the manuscript.

Reviewer #2; Comment 1

A major concern with the study is the uneven distribution of ages across the 4 public datasets. Most samples come from Onek1k and AIDA, which are enriched for an average age of ~40 (AIDA) and ~65 (Onek1k), which are the ages that appear to have the most DEFs. Further, since the individual datasets are clearly capturing either the first or second peak, it is plausible that the bimodal nature of the result is simply due to the uneven age distribution in the combined dataset.

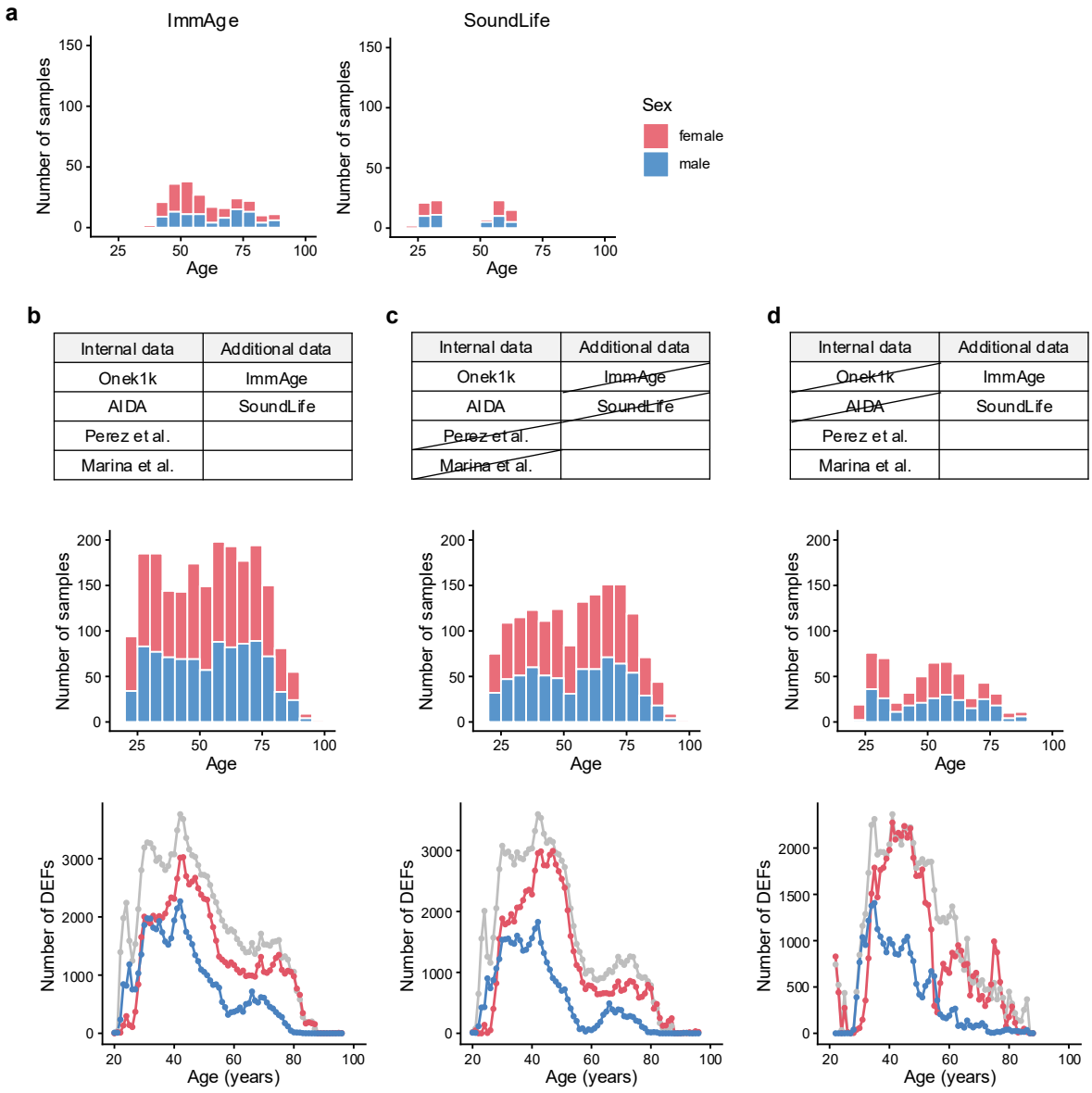
Response

This is a very important point and the Reviewer is correct by stating that the observed bimodal aging pattern could potentially arise from the uneven age distributions of the constituent datasets, particularly AIDA and OneK1K, which contribute a large proportion of samples and are enriched around approximately 40 and 65 years of age, respectively.

To evaluate whether this is the case, we opted for a step-by-step approach. We first expanded our analysis by incorporating two additional recently published PBMC datasets (ImmAge and SoundLife; see Methods lines 717-735; see Results lines 106-110). We then compared DE-SWAN results across three dataset combinations (Supplementary Figure 4 in the revised manuscript):

1. Original combined dataset plus the additional datasets
2. AIDA and OneK1K only
3. All datasets excluding AIDA and OneK1K, including the additional datasets.

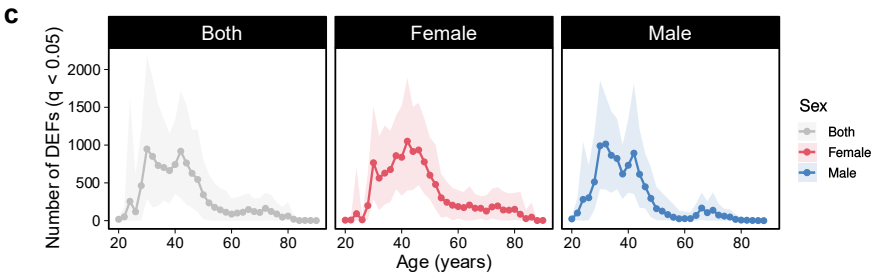
The critical comparison that answers directly the Reviewer's point is between (2) and (3). As expected, the AIDA- and Onek1k-only analysis reproduced the prominent peaks around approximately 40 and 65 years of age. Crucially, when AIDA and Onek1k were completely excluded, the bimodal pattern remained evident, with peaks observed at similar age ranges. The second peak remained particularly pronounced in the female-specific analysis.



Supplementary Figure 4. Reproducibility of DE-SWAN peaks using additional PBMC scRNA-seq datasets. **a**, Age and sex distribution of the ImmAge and SoundLife datasets, which were used as additional PBMC scRNA-seq data for validation. **b-d**, Summary of included datasets, their age and sex distributions, and DE-SWAN results under different dataset combinations: all internal and additional datasets (**b**); Onek1k and AIDA datasets only (**c**); and Perez et al., Marina et al., and additional datasets (**d**).

Related to this concern, we also performed a modified DE-SWAN analysis with downsampling, as described in our response to Reviewer #1, Comment 6, to directly assess whether the observed peaks could be driven by differences in sample density across age. Specifically, for

each age window tested (20-90 years, in 2-year increments), we randomly downsampled to 200 samples and repeated the analysis for 100 iterations. The bimodal pattern remained after downsampling, including the prominent first peak around 40 years of age and a smaller second peak around 65 years of age (most clearly observed in the male analysis; Extended Data Fig. 2c in the revised manuscript).



Extended Data Figure 2. DE-SWAN robustness analysis across parameters and cell types. **c**, Number of differentially expressed features (DEFs) across age identified by DE-SWAN using 100 random downsampling iterations, each using 200 samples.

Taken together, these complementary analyses demonstrate that the bimodal aging pattern is not solely driven by the age distributions of AIDA and Onek1k cohorts or by uneven sample density across age. The consistent observation of the two peaks across independent datasets and downsampling analyses supports the robustness and reproducibility of the identified aging transitions.

Reviewer #2; Comment 2

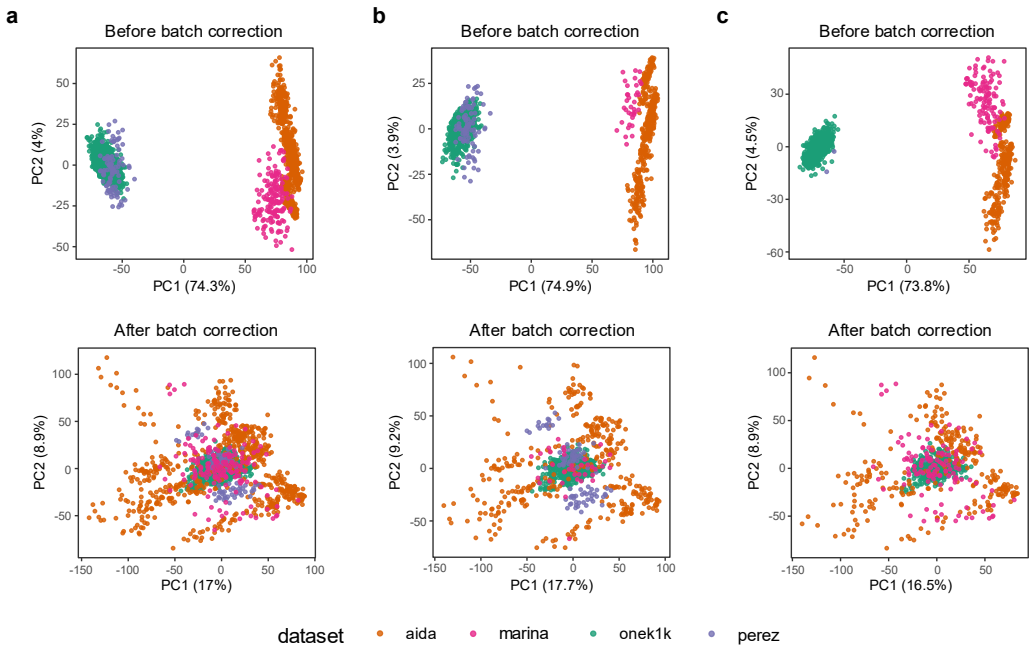
Related to the above point, although the authors include batch correction in the methods, there are no data (e.g. feature plots) showing batches before and after correction. This is important quality control, as it is possible that the different DE signatures in the two peaks are driven by differences in the AIDA and Onek1k datasets themselves.

Response

We thank the Reviewer for highlighting the importance of providing quality-control evidence for batch correction. This point overlaps with Reviewer #1's comments regarding dataset integration and batch effects (Reviewer #1, Comments #1-3).

As described in our response to Reviewer #1, dataset-associated variation was accounted for in the two downstream analyses. For DE-SWAN analysis, dataset identity was included as a covariate in the differential expression model, thereby directly controlling for dataset-specific effects during statistical testing. For clustering analysis, batch correction was performed on the merged pseudobulk expression matrices using limma's removeBatchEffect function prior to clustering and visualization.

To directly assess the effectiveness of batch correction, we have added PCA plots before and after correction in the revised manuscript (Supplementary Figure 4 in the revised manuscript). Prior to correction, samples exhibited partial separation by dataset, whereas this structure was substantially reduced after correction, indicating effective mitigation of dataset-associated variation.



Supplementary Figure 5. Batch effect correction for pseudobulk expression matrices. **a**-**c**, PCA plots before and after batch effect correction for both-sex (**a**), female (**b**), and male (**c**) pseudobulk expression matrices.

We additionally repeated the DE-SWAN analysis after applying removeBatchEffect to the pseudobulk expression matrix prior to differential expression analysis to evaluate whether the observed aging peaks were sensitive to the batch correction approach. The resulting age-associated patterns remained highly consistent with the original findings, including the presence and locations of the two major peaks and their associated cell-type signatures (Figure R1).

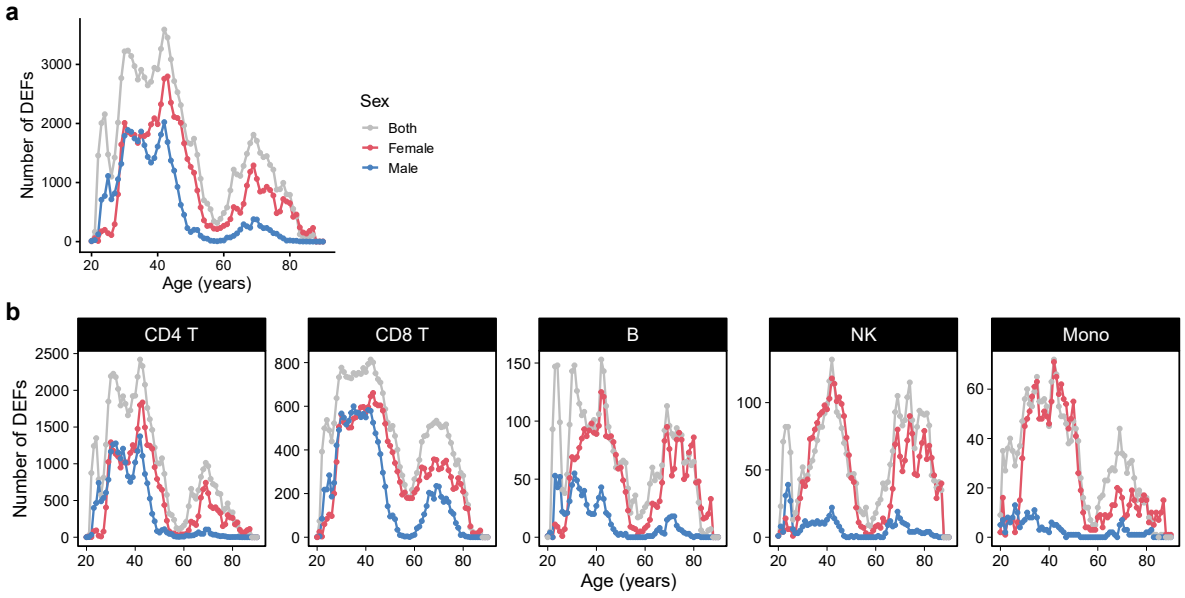


Figure R1. DE-SWAN results after applying removeBatchEffect. **a**-**b**, Number of differentially expressed features (DEFs, $q < 0.05$) across the age range identified by DE-SWAN algorithm after applying limma's removeBatchEffect, shown for all major PBMC cell types combined (**a**) and for each cell type separately (**b**).

Finally, the point that the two peaks may be driven by differences between the AIDA and OneK1K datasets is addressed above. Briefly, we showed that the bimodal pattern of aging remained evident even when AIDA and OneK1K were excluded from the analysis, with the two peaks observed at similar age ranges. Together, these analyses demonstrate that the observed bimodal aging pattern is not primarily driven by dataset-specific batch effects or the age distributions of the AIDA and OneK1K datasets.

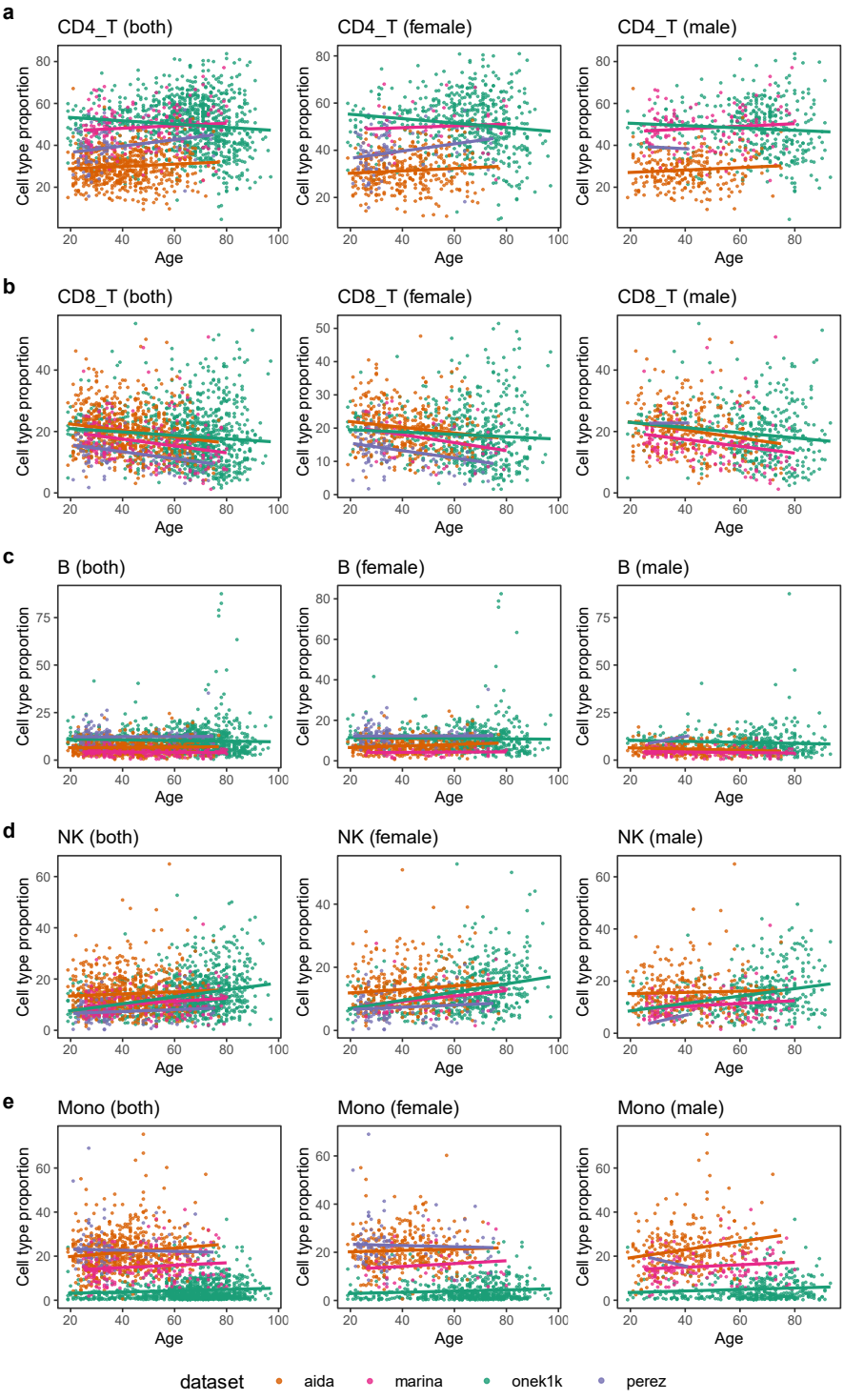
Reviewer #2; Comment 3

Data on the composition of each dataset should also be shown as this could bias the cell type analysis.

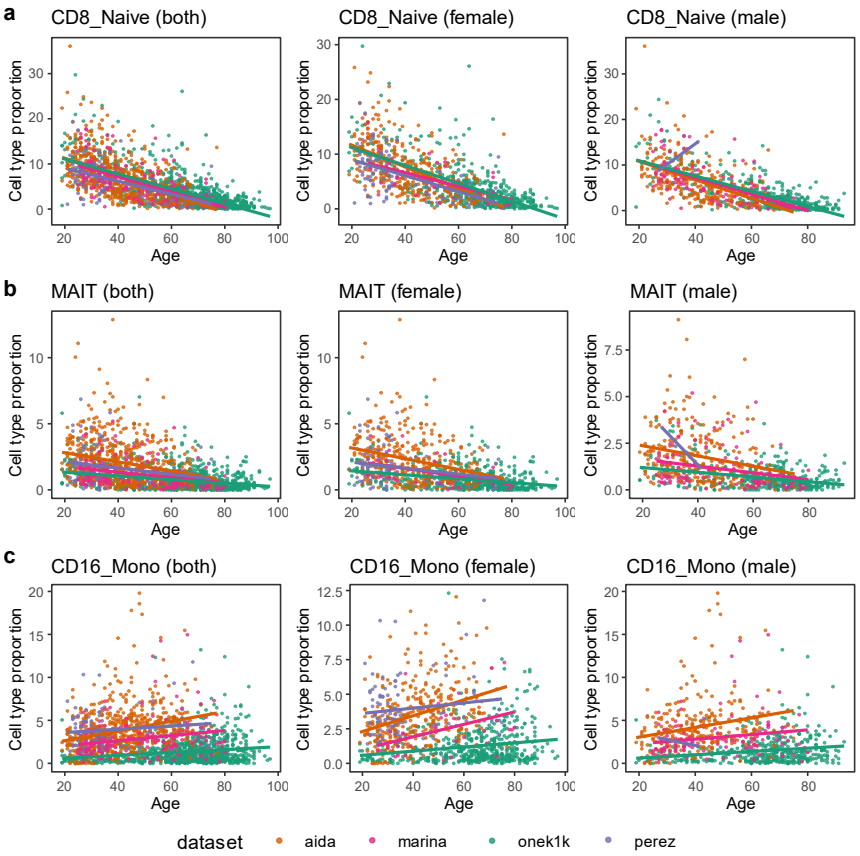
Response

We agree with the Reviewer. As described in our response to Reviewer #1 (Comment #4), we revised the cell type proportion analysis using the propeller framework (speckle R package), which is designed for differential cell type proportion analysis for scRNA-seq data, while accounting for biological replication and experimental covariates. In our analysis, dataset identity was included as a covariate to account for dataset-specific effects. The updated results are presented in the revised Table 1 (line 1472), and the relevant Results section has been revised accordingly (lines 83-86, lines 511-515).

To further assess the consistency of cell type proportion estimates across datasets, we generated dataset-stratified visualizations of cell type proportions. Supplementary Figures 2–3 in the revised manuscript show the proportions of the five major cell types (CD4 T cells, CD8 T cells, B cells, NK cells, and monocytes) as well as the cell subtypes highlighted in the manuscript (lines 511-515; CD8 naïve T cells, MAIT cells, and CD16 monocytes). Similar age-associated trends were observed across the independent cohorts, supporting that the reported findings are not driven by a single dataset.



Supplementary Figure 2. Age-associated changes in major cell type proportions across the four datasets. **a**-**e**, Cell type proportions plotted against age for CD4 T cells (**a**), CD8 T cells (**b**), B cells (**c**), NK cells (**d**), and monocytes (**e**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.



Supplementary Figure 3: Age-associated changes in detailed cell subtype proportions across the four datasets. **a**-**c**, Cell type proportions plotted against age for CD8 Naïve T cells (**a**), MAIT cells (**b**), and CD16 monocytes (**c**). Data are shown separately for all samples, female samples, and male samples. Points and regression lines are colored by each dataset.

Reviewer #2; Comment 4

The claim that the female-specific late-increase cluster represents a senescence-associated cellular state is not supported. Senescent cells are rare, even in aged individuals and their gene expression signatures are well defined. There is no actual analysis of senescence in this study and this claim (included in the abstract) is purely speculative.

Response

The Reviewer is correct and we agree that we do not have direct evidence on a senescent-associated cell state. As such, we no longer refer to this state as senescent but with the observed pathways (i.e. apoptosis-related, iron accumulation). According to the Reviewer's suggestion, we have modified the statements referring to senescence in the abstract (line 24), introduction (line 62), results (lines 264-266, 280), and discussion sections (lines 562-570).

Reviewer #2; Comment 5

Minor:

Extended Fig. 4 is not readable.

Response

We thank the Reviewer for pointing this out. We have revised the figure and replaced it with a higher resolution version to improve clarity.

Second Round of Revision

Reviewer #1 (Remarks to the Author)

The authors have sufficiently addressed my comments and concerns, and the manuscript has been significantly improved. I have only one minor suggestion: please ensure that all additional analyses performed in response to the reviewers' comments are incorporated into the revised manuscript. For example, Figure R2 does not appear to be included. Although these analyses may seem technical, including them is important for demonstrating the robustness of the presented results and will strengthen the manuscript.

Response

We thank the reviewer for their positive assessment and for suggesting that all additional analyses be included. We have now incorporated all additional analyses performed during the review process into the revised manuscript and supplementary materials. Specifically, Figure R1 has been added as Supplementary Fig. 5e-f, Figure R2 as Supplementary Fig. 8b, and Figure R3-4 as Supplementary Fig. 15. We have also revised the manuscript to reflect these additions (lines 104-105, 114-115, 617-619). We appreciate the reviewer's suggestion, as it strengthens the transparency and robustness of our findings.

Reviewer #2 (Remarks to the Author)

My concerns have been address and I am supportive of publication.

Response

We thank the reviewer for their time and constructive feedback throughout the review process, and for their support of publication.