nature portfolio

Peer Review File

Integrative epigenetics and transcriptomics identify aging genes in human blood

Corresponding Author: Professor Vadim Gladyshev

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)
In "Integrative epigenetics and transcriptomics identify aging genes in human blood," Moqri et al. present an integrative analysis of age-associated DNA methylation and RNA expression signatures. The authors can pinpoint roughly 100 genes robustly associated with aging by analyzing DNA methylation and RNA expression data from various cohorts. Importantly, enrichment analyses suggest that the identified genes may provide a better starting point for studying the molecular mechanisms of aging in the human blood. The study achieves this goal by generating an independent methylation data set to validate findings from previous, partially publicly available studies. Naturally, the generation of an independent RNA expression data set of the same cohort would have been an added benefit - although not strictly required, in my opinion, to substantiate the conclusions and claims of this study.

Conceptually, the manuscript introduces the term "multi-omic aging genes." While it may not be entirely new to seek genes exhibiting both transcriptional and regulatory changes during processes in time, the manuscript makes an essential case for analyzing multiple aging-associated features simultaneously. It demonstrates that this approach helps to identify biologically meaningful "aging genes."

The manuscript is well-written, and the presented data substantiates all conclusions and interpretations. Nevertheless, I have some suggestions that the authors may find helpful in revising their manuscript.

- The authors may like to discuss the advantages and limitations of DNAm arrays. For instance, arrays can measure only a fraction of all genomic CpGs, thus substantially limiting the detection of age-associated DNAm alterations. On the other hand, arrays may exhibit higher reproducibility, e.g., compared to RRBS or WGBS.

- The current manuscript implicitly goes beyond integrating DNAm and RNA expression already. Specifically, the authors show the enrichment of age-associated CpGs in regions affected by PRC2 binding. An increasing number of studies show that DNAm changes occurring during aging or in age-associated diseases are often associated with PRC2-associated silencing. Also, because PRC2 plays a role in bivalently marked genes that frequently play a role in development, I'd be curious whether the subset of multi-omic aging genes with PRC2 association enriches any more specific functions.

- The integration of DNAm and RNA expression data is straightforward but somewhat conservative. Specifically, the authors implicitly limited the regulatory region of an "aging transcript" by taking a region 1500bp upstream of the TSS. Naturally, this approach focuses on the promoter region but precludes the identification of many potentially relevant enhancer interactions. I wondered whether the authors have attempted to use epigenomic data, e.g., enhancer annotations for relevant blood cells, to detect such regulatory interactions.

- I am not sure whether this is out of scope, but it would be interesting to learn more about the regulatory similarities of the identified aging genes. For instance, whether the set of multi-omic aging genes can be associated with a set of upstream regulators or whether there is an enrichment of transcription factor binding sites, e.g., by integrating TF-ChIP data.

- The authors investigate the relationship of multi-omic aging genes and associated CpGs with mortality. In Figure 4c, Kaplan-Meier curves are shown for the top gene separately for the MGB-4K and GS data. I suggest the authors show the Kaplan-Meier curves for each gene/CpG in both data sets. For instance, adding the Kaplan-Meier plot for SATB1 (top gene

in MGB-4K) obtained from the GS data and the Kaplan-Meier plot for TCF7 (top gene in GS) obtained from the MGB-4K data. Since many of the CpGs/genes identified in MGB-4K and GS are identical, I propose showing Kaplan-Meier curves for all of them, e.g., in a supplement.

Minor recommendations:

- I may have missed the detailed description of the PRC2-associated analyses in the Methods section. Otherwise, I suggest elaborating on this a bit more.

- Although this information may be part of the code deposited on GitHub, I recommend explicitly stating which background model was used for GO analysis in the Methods section.

Reviewer #2

(Remarks to the Author)
The manuscript NCOMMS-24-55765-T entitled “Integrative epigenetics and transcriptomics identify aging genes in human blood” by Moqri et al. reports (i) the comparison, using six large cohorts, of the strongest correlations between DNAm levels and age with the strongest correlations between mRNA levels and age, replicating the previous observation that associations with the transcriptome are weaker; (ii) the observation that correlations between mRNA levels and age are weak for genes whose associated DNAm levels increase strongly with age, whereas correlations between DNAm and age are substantial for genes whose mRNA levels decrease strongly with age; (iii) evidence that genes showing an age-related increase in DNAm and an age-related decrease in mRNA (“multi-omic aging genes”) show more replicability across cohorts and are enriched in T-cell functions; (iv) evidence that aging CpGs are associated with the risk of dying with time and (v) aging CpGs remain associated with age when controlling for cellular composition.

The manuscript is well written and some results are of interest for the field. However, many aspects of the analyses are unclear and lack statistical support and justification. Please find below a list of comments and recommendations that should improve the manuscript.

Major comments
- Most statements in the manuscript lack statistical support. I invite the authors to systematically report p-values of statistical tests supporting these statements. Examples: “we found that genes with the greatest decrease in expression levels with aging consistently featured significantly higher DNAm correlation with age”; “we found considerably higher and more consistent correlations with age across the datasets”; “multi-omic aging genes to be particularly enriched for T cell-specific genes”.… Another salient example is “age-associated DNAm data are considerably more replicable across cohorts than transcriptomic data”: to support this claim, the authors should report the proportions of CpGs and genes that show a significant correlation with age both in the discovery and replication cohorts, and then test if this proportion is significantly higher for DNAm.
- The authors state that “multi-omic aging genes are strongly associated with clinically-relevant aging-related outcomes such as disease and mortality risk”. Actually, only DNAm levels at CpGs associated with aging genes are. Can the authors show volcano plots for expression levels of aging genes, as well as Peters et al.’s 1,497 genes? For comparison, I also suggest that the authors report volcano plots for the CpGs the most strongly associated with age, but that are not nearby aging transcripts. Finally, the authors are right that cellular composition can mediate the effects of age on DNAm or mRNA levels. I thus suggest that Cox regression models are adjusted on cellular composition.
- The authors observed that correlations between DNAm and age are substantial for genes whose mRNA levels decrease strongly with age. Accordingly, Peters et al. have shown that “1,248 of the 1,497 age-associated genes had ≥1 CpG site with a significant Sobel test”, implying that DNAm mediates most of the associations between age and gene expression at aging transcripts. This result should be acknowledged and discussed in length. Notably,
- It was previously demonstrated that CpG sites mediating age effects on gene expression are “less likely to reside in CpG islands or in promoters and more likely to be located in enhancers and insulators” (Peters et al., Nat Commun 2015). However, the authors of the current study defined “aging regulatory regions” as the 1,500 bases flanking the transcription start site of age-associated transcripts. Therefore, they may have missed most of the CpG sites that play a role in the regulation of genes whose expression is associated with age.
- I concur with the authors that aging genes could be markers of T cell differentiation, mirroring the age-related shift of T cells from naïve to memory states. The authors rightfully tested whether CpGs associated with aging genes are associated with age when controlling for the cell proportions estimated by Houseman et al.’s or Horvath’s deconvolution methods. However, none of these methods estimate the proportion of the cell types that are most likely to mediate age effects: memory T cells. I thus invite the authors to report associations of DNAm and RNA levels with age for aging genes controlling for the proportion of the 12 cell types estimated by EPIC IDOL-ext (Salas et al., Nat Commun 2022).
- The authors have used DNAm and RNA-seq from large cohorts of patients with age-related diseases (atherosclerosis, Parkinson). Therefore, disease status could mediate some of the strongest effects of age on DNAm levels or gene expression, whereas “aging genes” are supposed to be associated directly with age. Can the authors verify if their aging genes change when controlling for disease status (using multiple regression models)? Of note, the authors expect that “aging genes” could be used to “develop meaningful predictors for relevant aging-related outcomes”, but if these genes are

directly associated such aging-related outcomes, interpretation would very different.
- When the authors write “To validate our novel multi-omic aging genes, we aimed to perform an independent external validation using a novel cohort”, the reader does not know what are multi-omic aging genes, how they are defined exactly, their numbers... The authors cannot validate genes that are not yet identified.
- It is actually unclear whether and why the authors have used MGB only to define the multi-omic aging genes. It would make more sense to discover aging transcripts in the large RNA-seq cohorts (not only the 1,497 transcripts from Peters et al. 2015), discover aging CpGs and genes in the large DNam and RNA-seq cohorts and then validate these genes in MGB, whose sample size is more modest. Related to this, when are “validated aging genes” used in the study?
- The authors are right when they state that transcriptomic data show “higher noise, and susceptibility to batch effects” (Meyer et al., Aging Cell 2021). In the context of this study, were the RNA-seq data sets corrected for batch effects?
- Are non-“aging genes” with age-associated expression enriched for T cell-specific genes?
- Outlier values in DNAm and RNA-seq data can result in inflated correlations with age, particularly when the age distribution has long tails (e.g., 500FG, RA). Can the authors confirm their list of 106 aging genes with Spearman correlations?
- It would be very informative to identify further the features that characterize aging genes. Are these genes enriched in binding site for specific TFs?

Minor comments
- I suggest that the authors clarify what is the added value of identifying “aging genes” compared to identifying genes directly associated with age-related diseases.
- When the authors state their approach identifies “true aging-associated genes”, can they clarify what they mean?
- Could the authors speculate further on the causes of the less reproducible effects of age on gene expression across cohorts, compared to DNAm? The CD248 example suggests that age has a non-linear effect on its mRNA levels.
- Fig. 2: It could be informative to plot, for all genes, the correlation with age of gene expression against the correlation for associated DNAm.
- Please italicize gene names in the text and figures.
- In Fig 2c, it would be more informative to show results for PPMI instead of MESA 2010.
- Fig 3c, please replace “correlation” by “correlation”
- Table 1: Please add information on the tissue in which DNAm and RNA-seq were measured, and the percentage of diseased patients.
- Methods: Replace “Pearrson” by “Pearson”, and “18,85911 participants” by “18,859 participants”.
- Methods: Please describe in details the quality control filters applied on the new DNAm data.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The authors have addressed all my comments and suggestions in full.

Reviewer #2

(Remarks to the Author)
The authors have satisfactorily addressed most of my comments.
- My only remaining concern is the presence of batch effects within cohorts. Can the authors verify that first PCs of the RNA-seq and DNAm data do not reflect plate effects, beadchip effects or variation in sequencing coverage?

- L. 123: Please italicize gene names.
- Should the authors remove sentence in L. 130-131?
- Resolution of Suppl. Fig. 12 should be improved. Could the authors clarify which CpGs were selected for analysis shown in panel b?
- The Discussion is reported twice.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)
The authors have satisfactorily addressed my remaining comments.

Point-by-Point Responses

Reviewer #1:

In "Integrative epigenetics and transcriptomics identify aging genes in human blood," Moqri et al. present an integrative analysis of age-associated DNA methylation and RNA expression signatures. The authors can pinpoint roughly 100 genes robustly associated with aging by analyzing DNA methylation and RNA expression data from various cohorts. Importantly, enrichment analyses suggest that the identified genes may provide a better starting point for studying the molecular mechanisms of aging in the human blood. The study achieves this goal by generating an independent methylation data set to validate findings from previous, partially publicly available studies. Naturally, the generation of an independent RNA expression data set of the same cohort would have been an added benefit, although not strictly required, in my opinion, to substantiate the conclusions and claims of this study.

Conceptually, the manuscript introduces the term "multi-omic aging genes." While it may not be entirely new to seek genes exhibiting both transcriptional and regulatory changes during processes over time, the manuscript makes an essential case for analyzing multiple aging-associated features simultaneously. It demonstrates that this approach helps to identify biologically meaningful "aging genes."

The manuscript is well-written, and the presented data substantiates all conclusions and interpretations. Nevertheless, I have some suggestions that the authors may find helpful in revising their manuscript.

1. The authors may like to discuss the advantages and limitations of DNAm arrays. For instance, arrays can measure only a fraction of all genomic CpGs, thus substantially limiting the detection of age-associated DNA methylation alterations. On the other hand, arrays may exhibit higher reproducibility, e.g., compared to RRBS or WGBS.

Response: Thank you for the suggestion. DNA methylation arrays indeed measure only a subset of all genomic CpGs, which limits our ability to detect the full spectrum of age-associated DNAm alterations. Current generation arrays such as the Illumina MethylationEPIC v2.0 cover approximately 935,000 CpG sites, representing less than 5% of all CpG sites in the human genome. This selective coverage creates a potential bias in the identification of age-associated DNAm signatures.

However, arrays offer several advantages that make them particularly valuable for epigenetic aging studies. First, they demonstrate exceptional technical reproducibility with typical correlation coefficients exceeding 0.98 between technical replicates. Second, the

standardized protocols and established analysis pipelines significantly reduce technical variation between laboratories. Third, the relative cost-efficiency of arrays has enabled large-scale population studies that have been instrumental in defining robust epigenetic clocks.

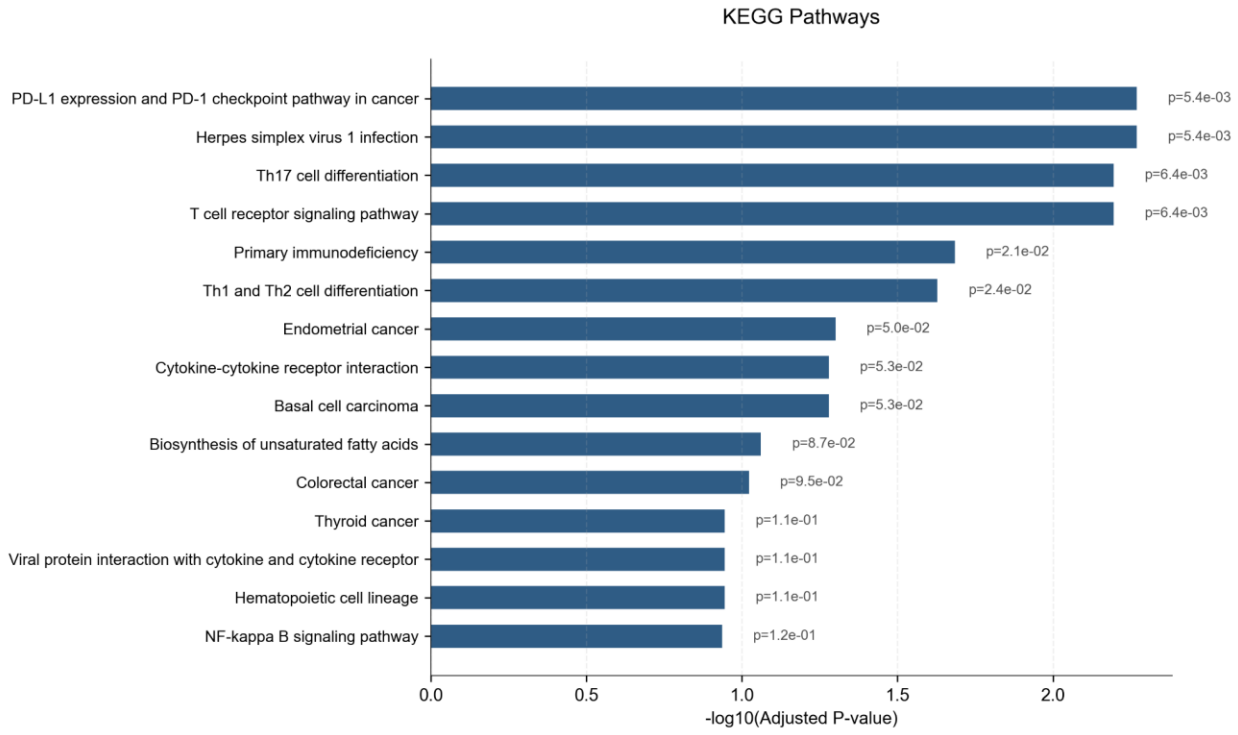
In contrast, whole-genome approaches such as WGBS (Whole Genome Bisulfite Sequencing) and RRBS (Reduced Representation Bisulfite Sequencing) offer more comprehensive coverage but face challenges in reproducibility and data processing. WGBS provides the most comprehensive assessment of methylation status but requires significant sequencing depth to achieve accuracy comparable to arrays, substantially increasing costs for large-scale studies. RRBS offers a middle ground by enriching for CpG-rich regions but suffers from higher technical variability compared to arrays.

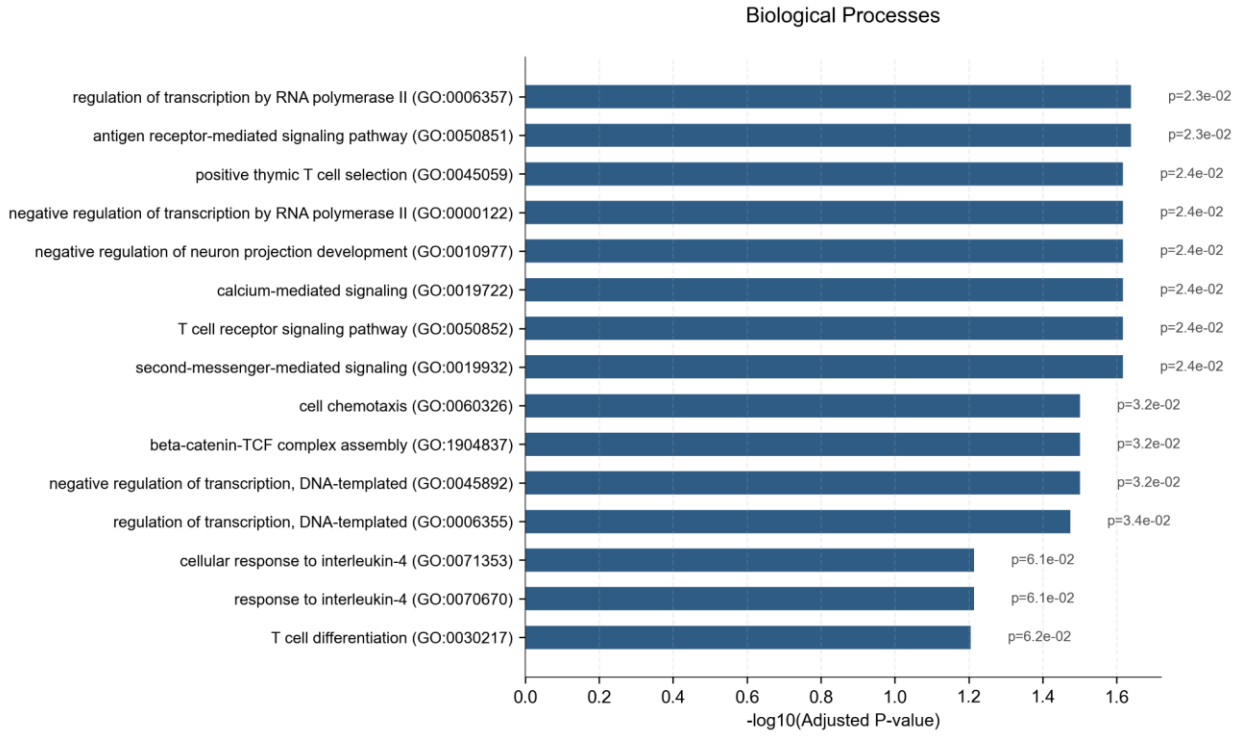
As suggested by this reviewer, we have added a brief section to our discussion, as follows:

> *"While the DNA methylation array technology applied measures only a subset of genomic CpG (approximately 935,000 CpG sites, representing 5%) in contrast to sequencing approaches, such as whole-genome bisulfite sequencing (WGBS), this technology offers several advantages for epigenetic studies, including technical reproducibility, standardized protocols and established analysis pipelines that significantly reduce technical variation between laboratories, and cost-efficiency"*

2. The current manuscript implicitly goes beyond integrating DNAm and RNA expression already. Specifically, the authors show the enrichment of age-associated CpGs in regions affected by PRC2 binding. An increasing number of studies show that DNAm changes occurring during aging or in age-associated diseases are often associated with PRC2-associated silencing. Also, because PRC2 plays a role in bivalently marked genes that frequently play a role in development, I'd be curious whether the subset of multi-omic aging genes with PRC2 association enriches any more specific functions.

Response: We appreciate the reviewer's insightful observation regarding PRC2 binding. Our enrichment analysis provides strong support for this connection, confirming our previous findings (Moqri et al. 2024, see figures below). Specifically, our ChEA transcription factor analysis revealed significant enrichment of PRC2-associated factors ($p < 0.05$) among our multi-omic aging genes. This aligns with the growing literature linking age-associated DNA methylation changes to PRC2-mediated regulation. Furthermore, our GO Biological Process analysis identified significant enrichment of developmental processes, consistent with the bivalent marking characteristic of PRC2 target genes. These findings strengthen the mechanistic link between aging-associated epigenetic changes and developmental regulatory networks.



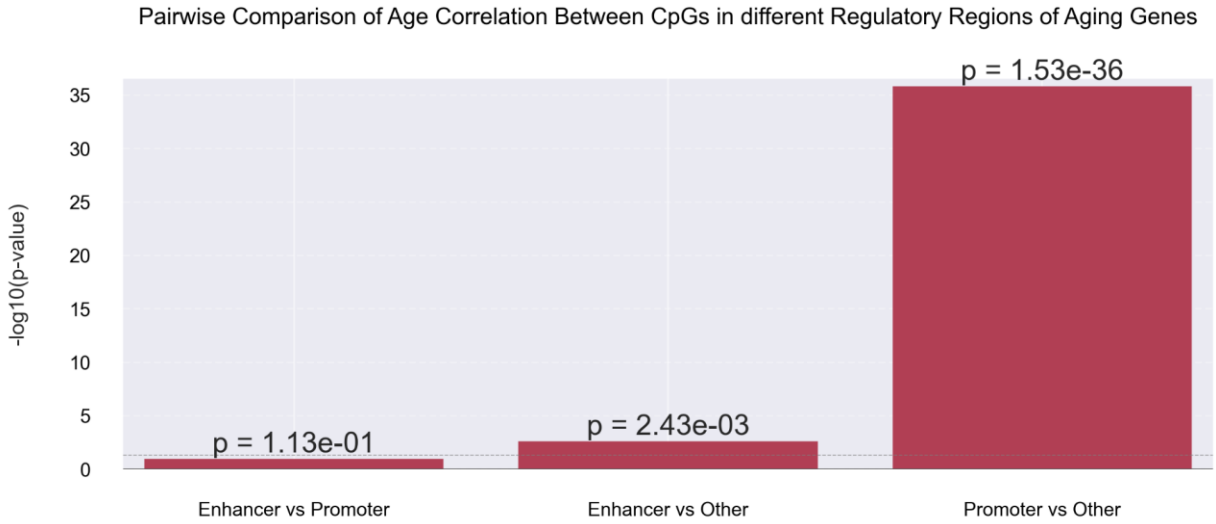


3. The integration of DNAm and RNA expression data is straightforward but somewhat conservative. Specifically, the authors implicitly limited the regulatory region of an "aging transcript" by taking a region 1500bp upstream of the TSS. Naturally, this approach focuses on the promoter region but precludes the identification of many potentially relevant enhancer

interactions. I wondered whether the authors have attempted to use epigenomic data, e.g., enhancer annotations for relevant blood cells, to detect such regulatory interactions.

Response: We appreciate the reviewer's comment regarding CpG sites associated with age-related gene expression. We have now assessed this using a systematic comparison of age correlation across different regulatory regions in aging-associated genes.

As shown in our analysis (added to the Supplement), we found that CpG sites in promoter regions demonstrate dramatically stronger age correlations compared to other genomic regions ($p = 1.53e-36$). The difference between enhancer regions and promoters was not statistically significant ($p = 1.13e-01$), while enhancers showed modestly stronger correlations than other regions ($p = 2.43e-03$). Please see the response to Reviewer 2 question 4 for more details.

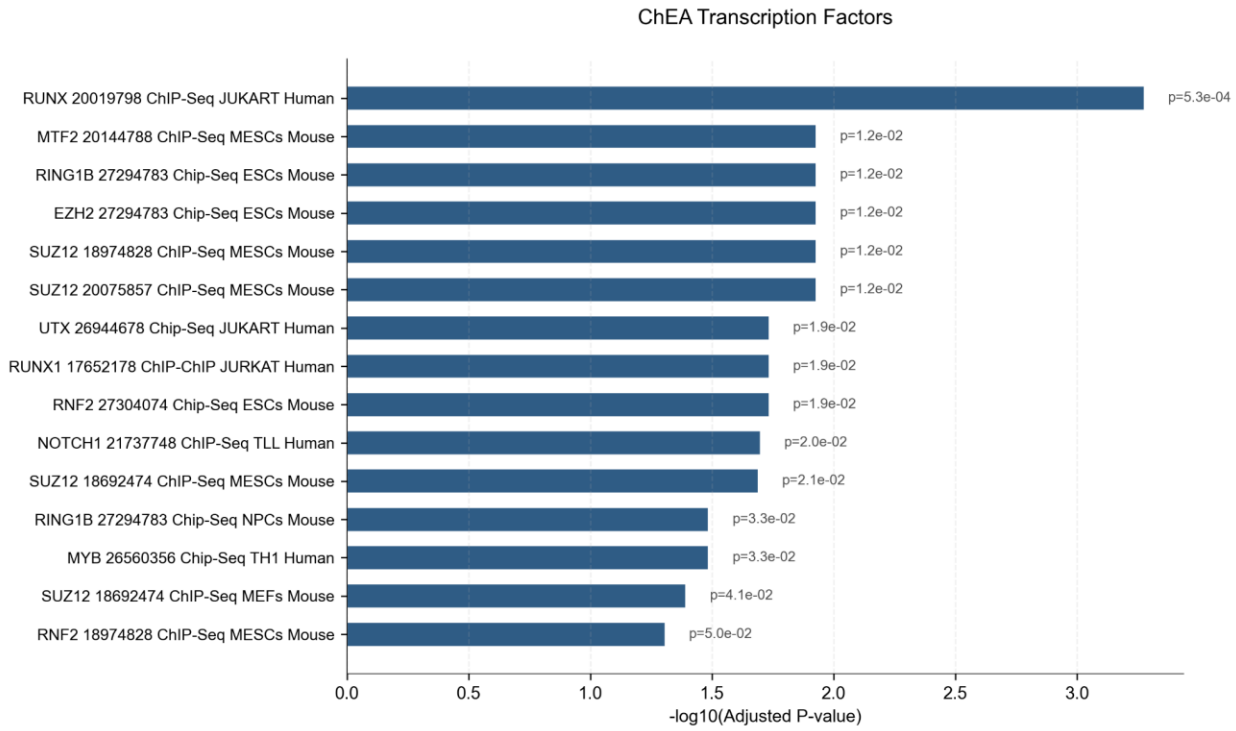


4. I am not sure whether this is out of scope, but it would be interesting to learn more about the regulatory similarities of the identified aging genes. For instance, whether the set of multi-omic aging genes can be associated with a set of upstream regulators or whether there is an enrichment of transcription factor binding sites, e.g., by integrating TF-ChIP data.

Response: We thank the reviewer for this excellent suggestion. We have conducted comprehensive transcription factor enrichment analyses using integration of ChIP-seq data through the ENCODE and ChEA databases revealed significant enrichment for specific transcription factor binding (15 significant TF associations, $p < 0.05$). This analysis identified key upstream regulators coordinating the expression of our multi-omic aging genes, notably many of which are associated with the PRC2 complex (*EZH2*, *SUZ12*) or immune/haematopoietic cell development (e.g., *RUNX*).

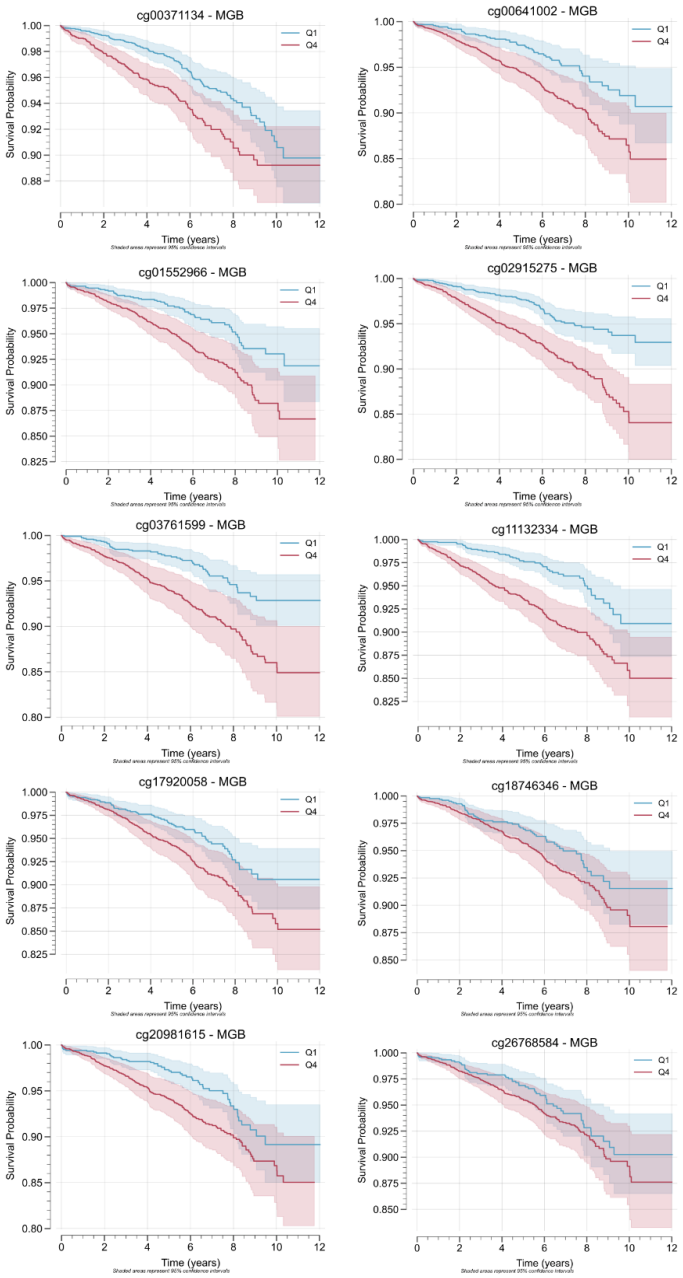
These findings reveal a coordinated regulatory network underlying age-associated epigenetic and transcriptional changes, rather than isolated alterations. This comprehensive regulatory

analysis strengthens our understanding of the molecular mechanisms driving aging-associated genomic changes and provides new insights into potential therapeutic targets for age-related conditions.

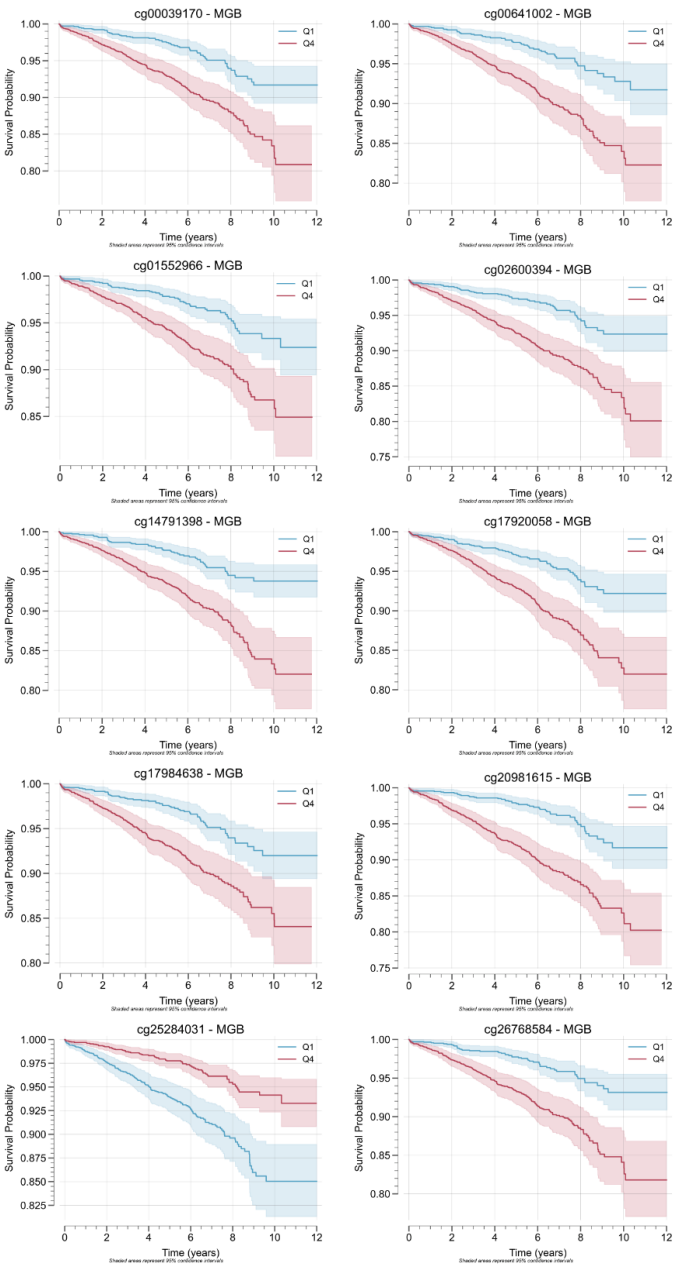


5. The authors investigate the relationship of multi-omic aging genes and associated CpGs with mortality. In Figure 4c, Kaplan-Meier curves are shown for the top gene separately for the MGB-4K and GS data. I suggest the authors show the Kaplan-Meier curves for each gene/CpG in both data sets. For instance, adding the Kaplan-Meier plot for SATB1 (top gene in MGB-4K) obtained from the GS data and the Kaplan-Meier plot for TCF7 (top gene in GS) obtained from the MGB-4K data. Since many of the CpGs/genes identified in MGB-4K and GS are identical, I propose showing Kaplan-Meier curves for all of them, e.g., in a supplement.

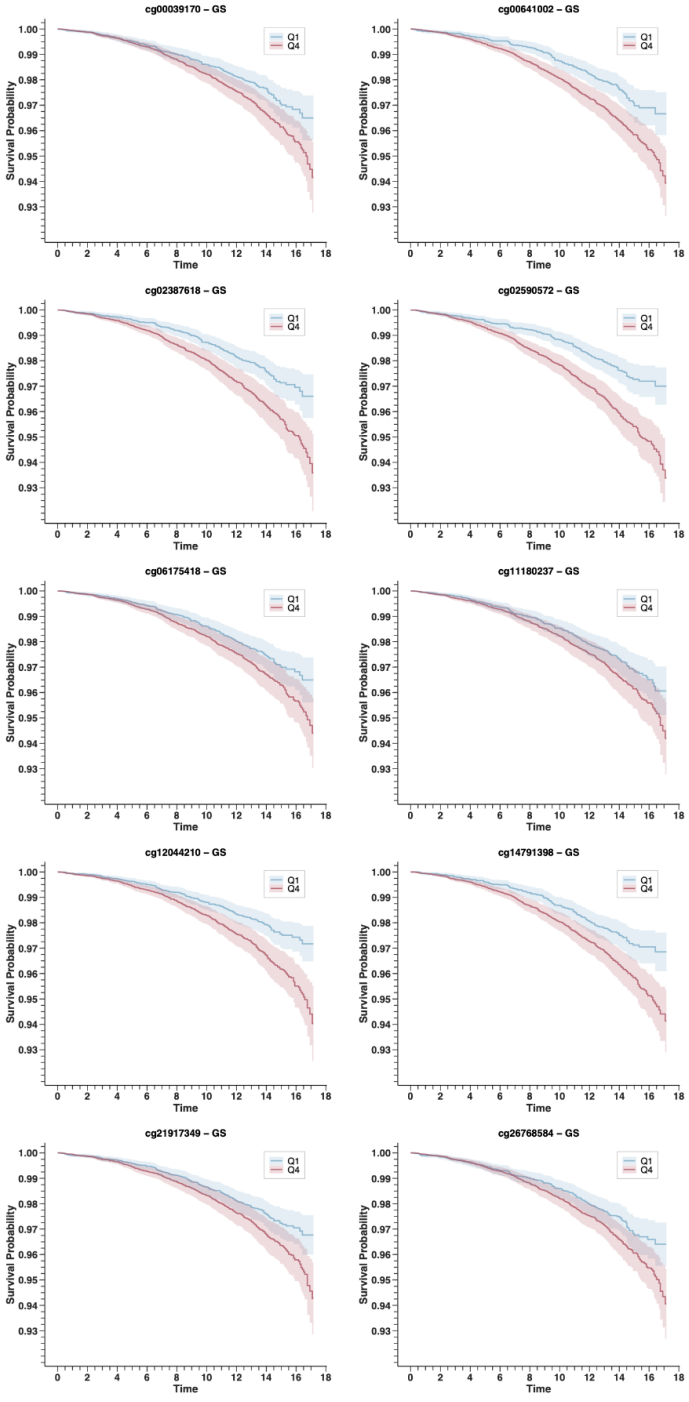
Response: We thank the reviewer for this suggestion. We have extended our KM analyses to include all the top CpG sites and included the results in the supplement, and copied below.



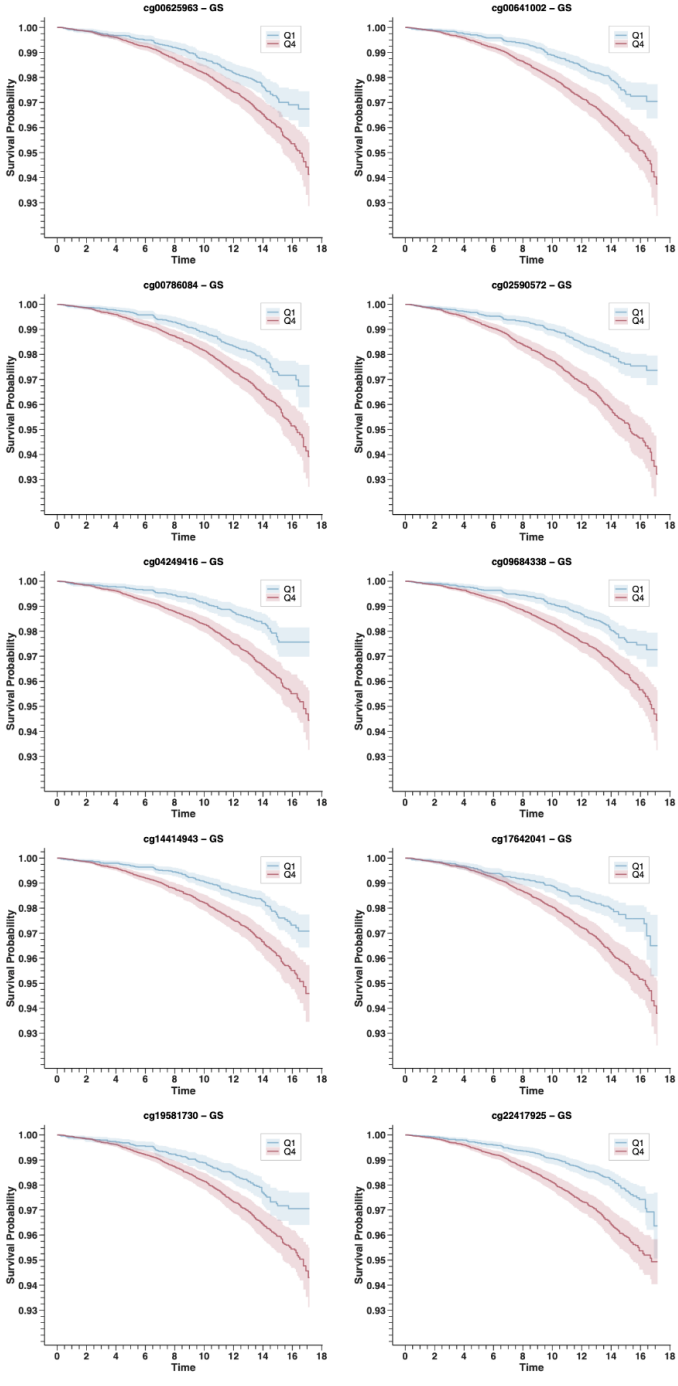
MGB Top CpGs Survival Plots, Adjusted for Cell composition



MGB Top CpGs Survival Plots, Non-Adjusted for Cell composition



GS Cohort Top CpGs Survival Plots, Adjusted for Cell composition



GS Cohort Top CpGs Survival Plots, Adjusted (no cell composition)

Minor recommendations:

- I may have missed the detailed description of the PRC2-associated analyses in the Methods section. Otherwise, I suggest elaborating on this a bit more.

Response: We thank the reviewer for this suggestion. We have updated the figure legend to refer the readers to the Method section for more details on the PRC2-associated analysis and added the following details under the Methods:

> *"Following Moqri et al. 2024, PRC2 binding regions were identified using EZH2 binding regions based on ChIP-Seq data from ENCODE Consortium (ENCFF109KCQ) and CpGs within these regions are marked as PRC2-associated sites"*

- Although this information may be part of the code deposited on GitHub, I recommend explicitly stating which background model was used for GO analysis in the Methods.

Response: We thank the reviewer for this suggestion. We have updated the text to include the following details about the GO analyses under the Methods:

> *"Functional enrichment of genes associated with aging across modalities was performed using Fisher's exact test and gene terms from GO Biological Process, using all available genes in the ontology with clusterProfiler package in R. Redundant functions with high semantic similarity were filtered out with simplify function"*

Reviewer #2:

The manuscript NCOMMS-24-55765-T entitled “Integrative epigenetics and transcriptomics identify aging genes in human blood” by Moqri et al. reports (i) the comparison, using six large cohorts, of the strongest correlations between DNAm levels and age with the strongest correlations between mRNA levels and age, replicating the previous observation that associations with the transcriptome are weaker; (ii) the observation that correlations between mRNA levels and age are weak for genes whose associated DNAm levels increase strongly with age, whereas correlations between DNAm and age are substantial for genes whose mRNA levels decrease strongly with age; (iii) evidence that genes showing an age-related increase in DNAm and an age-related decrease in mRNA (“multi-omic aging genes”) show more replicability across cohorts and are enriched in T-cell functions; (iv) evidence that aging CpGs are associated with the risk of dying with time and (v) aging CpGs remain associated with age when controlling for cellular composition.

The manuscript is well written and some results are of interest for the field. However, many aspects of the analyses are unclear and lack statistical support and justification. Please find below a list of comments and recommendations that should improve the manuscript.

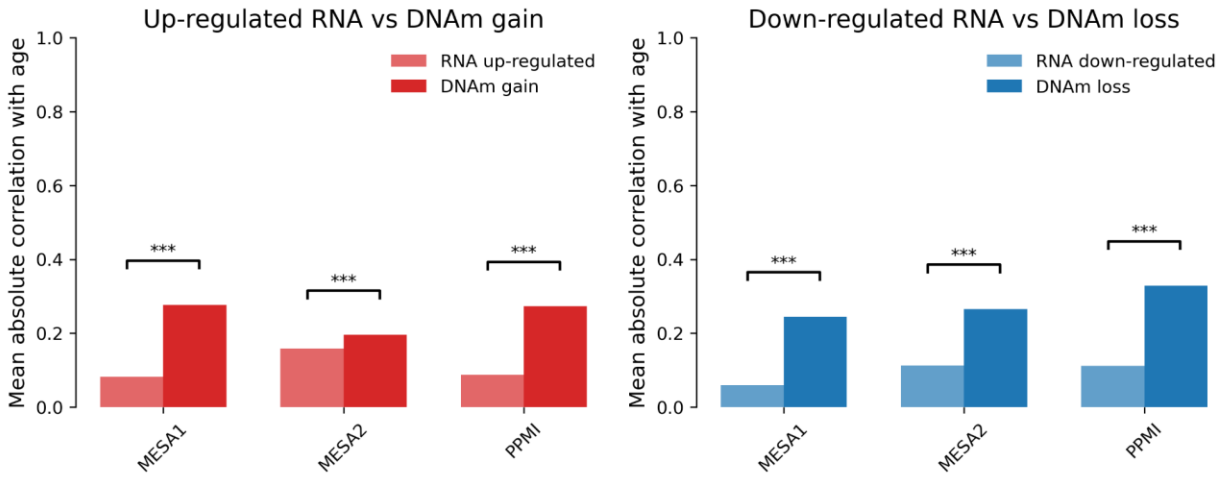
Major comments

1. Most statements in the manuscript lack statistical support. I invite the authors to systematically report p-values of statistical tests supporting these statements. Examples: “we found that genes with the greatest decrease in expression levels with aging consistently featured significantly higher DNAm correlation with age”; “we found considerably higher and more consistent correlations with age across the datasets”; “multi-omic aging genes to be particularly enriched for T cell-specific genes”…. Another salient example is “age-associated DNAm data are considerably more replicable across cohorts than transcriptomic data”: to support this claim, the authors should report the proportions of CpGs and genes that show a significant correlation with age both in the discovery and replication cohorts, and then test if this proportion is significantly higher for DNAm.

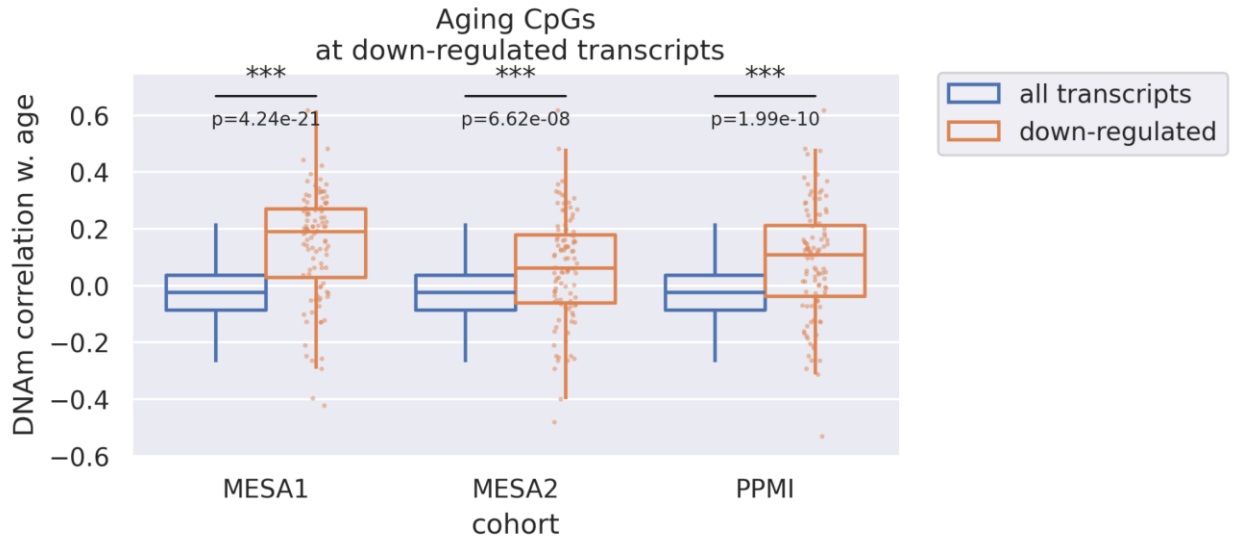
Response: We thank the reviewer for their valuable feedback regarding the need for statistical support for our claims. We have now conducted comprehensive statistical tests throughout our analysis and have updated the figures to include the relevant p-values and significance indicators.

We have analyzed the replicability of DNAm and transcriptomic data across cohorts using formal statistical testing. Our analysis confirms that age-associated DNAm data are significantly more replicable as aging CpGs showing higher correlations with age across

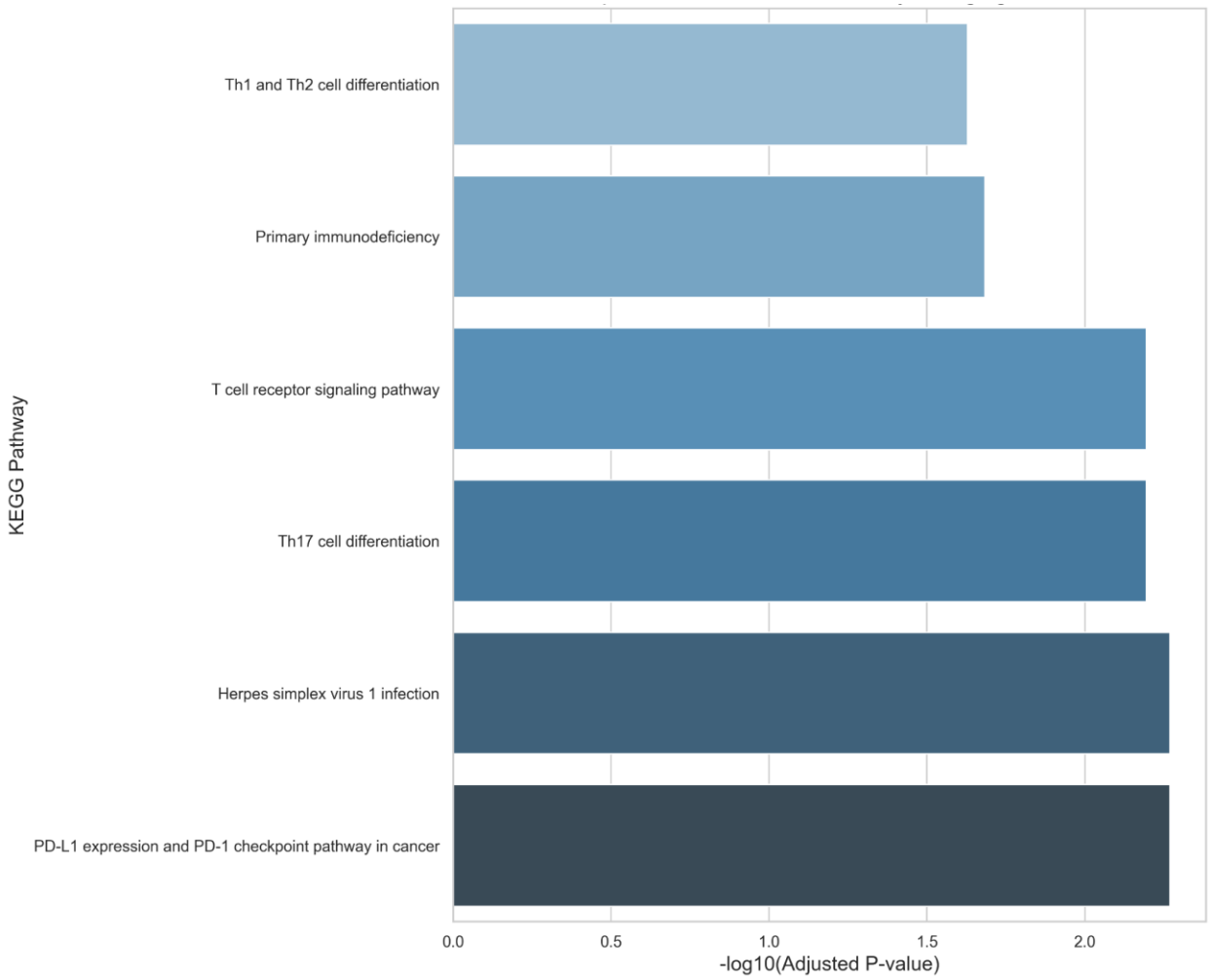
datasets compared to aging transcripts.



We have also validated the claim that "genes with the greatest decrease in expression levels with aging consistently featured significantly higher DNAm correlation with age." Statistical testing across all three cohorts (MESA1, MESA2, and PPMI) demonstrated highly significant differences (Wilcoxon rank-sum test, p=4.24e-21 for MESA1, p=6.62e-08 for MESA2, and p=1.99e-10 for PPMI), as indicated by the p-values displayed directly in figure 3f.



Regarding our claim about multi-omic aging genes being "particularly enriched for T cell-specific genes," we have conducted formal enrichment analysis as shown in the graph below. These results demonstrate significant enrichment for T-cell specific functions, including T cell receptor signaling pathway (adjusted p-value < 0.01) and T cell differentiation pathways (Th1/Th2 and Th17 cell differentiation, adjusted p-values < 0.05).



Throughout our study, we have systematically applied appropriate statistical testing to support our claims. We have updated all figures with statistical tests to include p-value annotations. We believe these statistical tests address the reviewer's concerns and provide robust support for the key claims in our manuscript.

2. The authors state that “multi-omic aging genes are strongly associated with clinically-relevant aging-related outcomes such as disease and mortality risk”. Actually, only DNAm levels at CpGs associated with aging genes are. Can the authors show volcano plots for expression levels of aging genes, as well as Peters et al.’s 1,497 genes? For comparison, I also suggest that the authors report volcano plots for the CpGs the most strongly associated with age, but that are not nearby aging transcripts. Finally, the authors are right that cellular composition can mediate the effects of age on DNAm or mRNA levels. I thus suggest that Cox regression models are adjusted on cellular composition.

Response: We appreciate the reviewer's comment regarding the association between multi-omic aging genes and aging-related outcomes. We would like to clarify our approach and provide additional evidence supporting our conclusions.

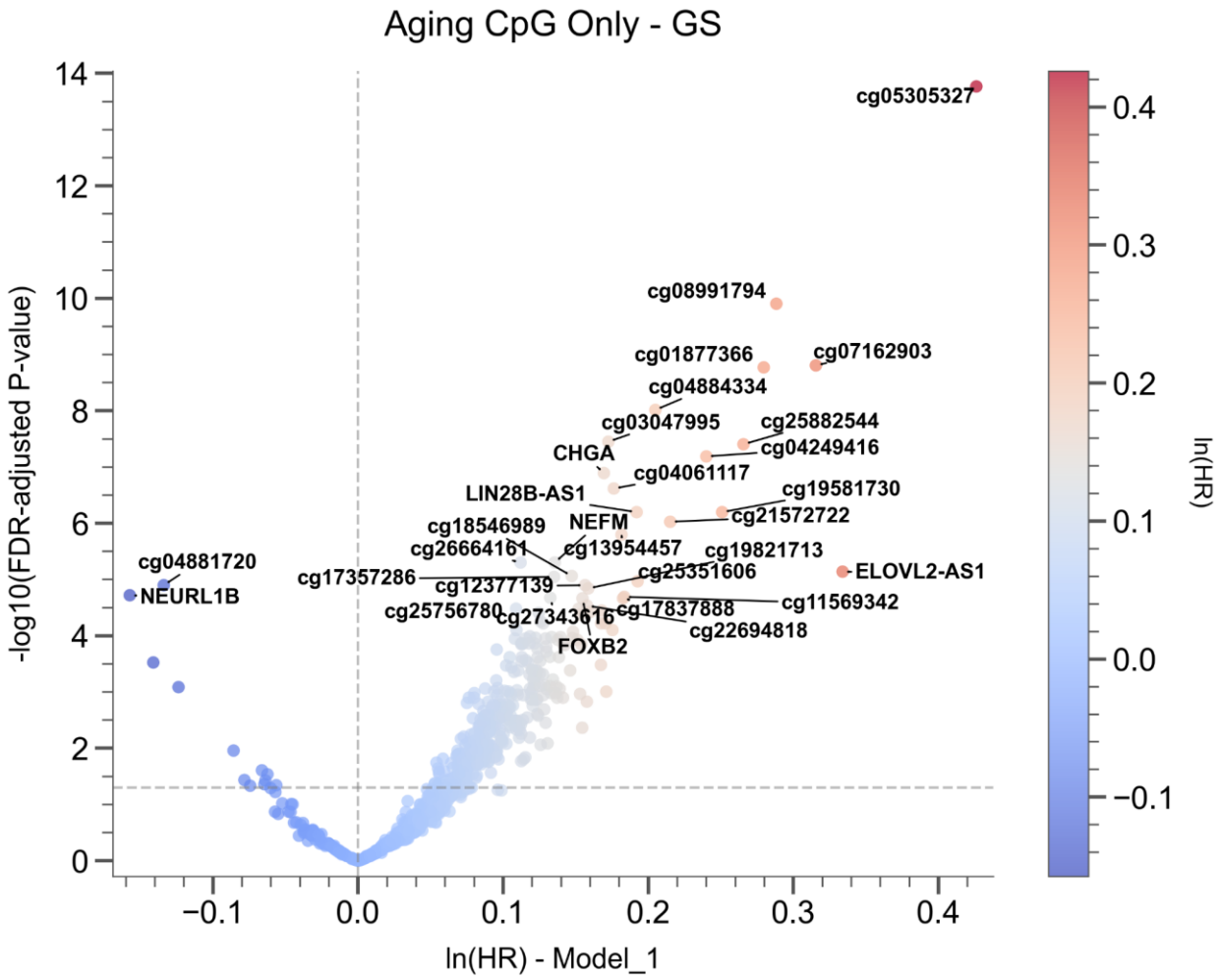
As the reviewer correctly notes, it is the DNAm levels at CpGs associated with aging genes that show associations with clinical outcomes. We do not have access to gene expression data in the MGB and GS cohorts used for outcome analyses, which prevents direct comparison of expression-based associations. Instead, we conducted an alternative analysis to address this question.

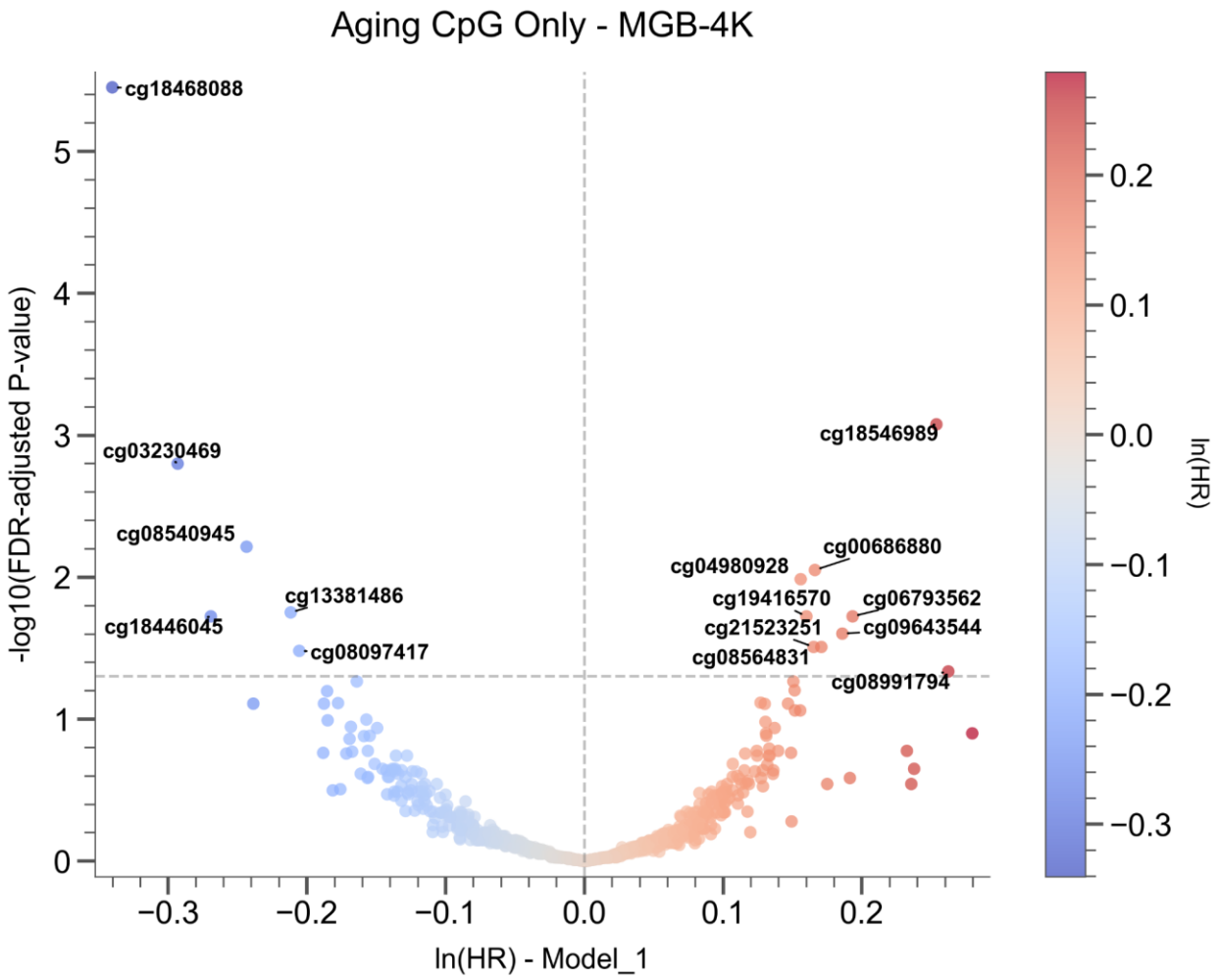
To address the reviewer's concern, we selected the top 1,000 CpGs from MGB500 that showed positive age correlation (DNAm gain), without pre-filtering based on association with aging-related genes. We then generated volcano plots for mortality prediction in both GS and MGB4K cohorts. All Cox regression models were adjusted for age and 12 different cell type compositions to account for the cellular heterogeneity concerns raised by the reviewer.

The volcano plots demonstrate that CpGs selected based on our multi-omic approach show stronger and more consistent associations with mortality across both cohorts compared to CpGs selected based solely on age correlation. The effect sizes for the non-gene-informed CpGs were notably smaller, with fewer sites reaching statistical significance after multiple testing correction.

This cross-cohort validation supports our conclusion that integrating transcriptomic information with epigenetic data identifies more clinically relevant CpG sites. The adjustment for cellular composition in our models further ensures that these associations are not merely reflecting age-related changes in blood cell proportions.

These findings reinforce our central claim that the multi-omic approach yields more robust biomarkers of biological aging with stronger links to health outcomes than approaches using either data type in isolation.


Aging CpG Only - GS
-log10(FDR-adjusted P-value)
ln(HR) - Model_1
ln(HR)
cg05305327
cg08991794
cg01877366
cg07162903
cg04884334
cg03047995
cg25882544
CHGA
cg04249416
cg04061117
LIN28B-AS1
cg19581730
cg21572722
cg18546989
NEFM
cg26664161
cg13954457
cg19821713
cg04881720
cg17357286
cg25351606
ELOVL2-AS1
NEURL1B
cg12377139
cg25756780
cg27343616
cg17837888
cg11569342
cg22694818
FOXB2



3. The authors observed that correlations between DNAm and age are substantial for genes whose mRNA levels decrease strongly with age. Accordingly, Peters et al. have shown that “1,248 of the 1,497 age-associated genes had ≥1 CpG site with a significant Sobel test”, implying that DNAm mediates most of the associations between age and gene expression at aging transcripts. This result should be acknowledged and discussed in length.

Response: Thank you. We have now inserted a paragraph in the discussion section:

> *“Our observation that genes with strong age-related decreases in expression also tend to show substantial age-associated DNAm changes is consistent with some prior findings. Peters et al. (Nat Commun 2015) suggested that DNAm mediates the majority of associations between age and gene expression, with 1,248 of 1,497 age-associated genes showing at least one CpG site with a significant Sobel test. However, there are notable differences between our analyses. While they reported that such mediating*

CpGs were more likely to reside in enhancers and less likely in promoters which is in contrast with our own findings, our analysis moreover focused specifically on CpG sites showing robust methylation gain with age, a process known to preferentially affect promoter regions. Additionally, our study sought to identify "multi-omic aging genes" where both expression and DNAm shift with age, rather than test mediation. Thus, while the mechanistic focus differs, our findings are complementary and highlight the complexity of DNAm regulation across genomic contexts during aging."

4. It was previously demonstrated that CpG sites mediating age effects on gene expression are "less likely to reside in CpG islands or in promoters and more likely to be located in enhancers and insulators" (Peters et al., Nat Commun 2015). However, the authors of the current study defined "aging regulatory regions" as the 1,500 bases flanking the transcription start site of age-associated transcripts. Therefore, they may have missed most of the CpG sites that play a role in the regulation of genes whose expression is associated with age.

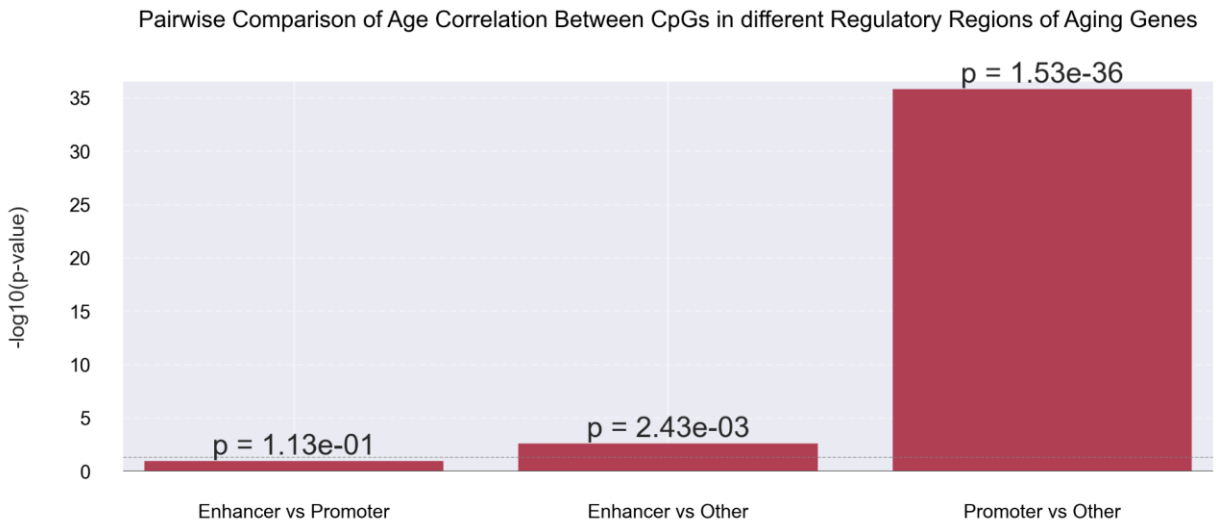
Response: We appreciate the reviewer's comment regarding CpG sites associated with age-related gene expression. We have now assessed this point through a systematic comparison of age correlation across different regulatory regions in aging-associated genes.

As shown in our analysis (added to the supplement, and figure copied below), we found that CpG sites in promoter regions demonstrate dramatically stronger age correlations compared to other genomic regions ($p = 1.53e-36$). The difference between enhancer regions and promoters was not statistically significant ($p = 1.13e-01$), while enhancers showed modestly stronger correlations than other regions ($p = 2.43e-03$).

This finding complements but also somewhat differs from Peters et al. (Nat Commun 2015), who reported that CpG sites mediating age effects were "less likely to reside in CpG islands or in promoters and more likely to be located in enhancers and insulators." It is important to note that our study's focus is fundamentally different. While Peters et al. identified CpG sites that mediate the effect of age on gene expression through Sobel mediation tests, our study specifically aimed to identify robust "multi-omic aging genes" by integrating epigenetic and transcriptomic age-related changes.

Moreover, we specifically focused on age-related methylation gain, which has been shown in numerous studies to preferentially affect promoter regions. The strong statistical significance we observed ($p = 1.53e-36$) validates our methodological choice to focus on the 1,500 bases flanking transcription start sites as "aging regulatory regions."

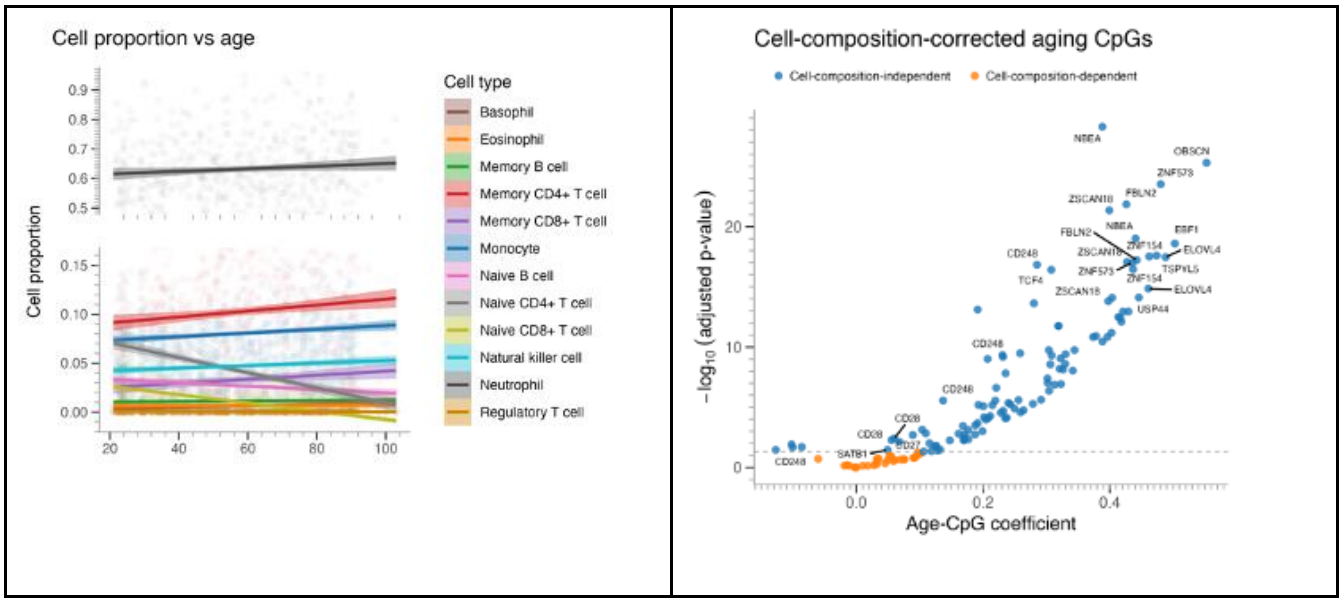
This approach has yielded a set of multi-omic aging genes that demonstrate greater replicability across diverse populations and stronger associations with aging-related outcomes compared to genes identified using either epigenetic or transcriptomic data alone, as detailed in our manuscript.



5. I concur with the authors that aging genes could be markers of T cell differentiation, mirroring the age-related shift of T cells from naïve to memory states. The authors rightfully tested whether CpGs associated with aging genes are associated with age when controlling for the cell proportions estimated by Houseman et al.'s or Horvath's deconvolution methods. However, none of these methods estimate the proportion of the cell types that are most likely to mediate age effects: memory T cells. I thus invite the authors to report associations of DNAm and RNA levels with age for aging genes controlling for the proportion of the 12 cell types estimated by EPIC IDOL-ext (Salas et al., Nat Commun 2022).

Response. Thank you for this excellent point. In collaboration with the developer of IDOL (who has now joined as a co-author of this work), we have included the results for 12 cell types in the revision which is in line with our findings with 6 cell types. We have included these figures in the supplement and updated the text accordingly as follows:

> *"Significant associations of methylation levels at these CpGs with age remained after the adjustment (Fig. 5e), indicating that age-related changes in these CpGs are independent of changes in cell composition. The results are consistent when adjusting for cell type compositions using 12 cell-type deconvolution (see Supplement)"*



6. The authors have used DNAm and RNA-seq from large cohorts of patients with age-related diseases (atherosclerosis, Parkinson). Therefore, disease status could mediate some of the strongest effects of age on DNAm levels or gene expression, whereas “aging genes” are supposed to be associated directly with age. Can the authors verify if their aging genes change when controlling for disease status (using multiple regression models)? Of note, the authors expect that “aging genes” could be used to “develop meaningful predictors for relevant aging-related outcomes”, but if these genes are directly associated such aging-related outcomes, interpretation would very different.

Response: This is an important point and we fully agree with this reviewer that disease states can modulate epigenetic features.

All of our integrative aging genes are selected from the initial set of transcriptomic aging genes identified using meta analyses of large diverse generally healthy or community-based cohorts by Peters et al. 2015. In addition, we have selected our epigenetic datasets from large cohorts of mostly healthy individuals and excluded diseased subjects from these cohorts. In particular our reference DNA methylation dataset, MGB500, is selected from healthy controls donors in the MGB Biobank, MESA participants were recruited from a generally healthy population (without any cardiovascular diseases), and Generation Scotland participants are generally healthier than the population (Navrady et al. 2017). We have now also included additional details about the cohorts (in addition to the existing description of MGB4k and GS) and clarified the exclusion of diseased samples from specific cohorts, including the following:

"MGB500 samples were selected from healthy control donors of diverse ages from the MGB Biobank. The donors were recruited from a major metropolitan academic medical center, were roughly balanced between male and female, and were generally representative of the racial/ethnic distribution of the local area (Fig. 3a).

MESA cohort is a population-based sample of an initial 6,814 men and women aged 45-84 without known cardiovascular disease. Thirty-eight percent of the recruited participants were White, 28 percent African American, 22 percent Hispanic, and 12 percent Asian, predominantly of Chinese descent. Participants were recruited from six field centers across the United States: Wake Forest University, Columbia University, Johns Hopkins University, University of Minnesota, Northwestern University and University of California - Los Angeles. (mesa-nhlbi.org).

Public DNAm and RNA-seq data were selected among commonly used datasets used in epigenetic or transcriptomic studies of aging from the generally healthy populations, with the exception of PPMI which was included to increase our multi-omic data. Diseased participants from the GC6 cohort were excluded from the analyses."

7. When the authors write "To validate our novel multi-omic aging genes, we aimed to perform an independent external validation using a novel cohort", the reader does not know what are multi-omic aging genes, how they are defined exactly, their numbers... The authors cannot validate genes that are not yet identified.

Response: This is an important clarification. Thank you for noting. We have updated the text to elaborate this as follows.

"To demonstrate that our integrative analyses and the identified epigenetic-transcriptomic aging genes are replicable, we performed an independent external validation using a novel cohort we specifically designed for omic analysis of aging biomarkers across lifespan. We generated DNAm profiles for 500 individuals of diverse ages from the Mass General Brigham (MGB) Biobank using the Illumina Infinium MethylationEPIC v2.0 array, which covers over 935,000 CpG sites enriched for regulatory regions and has been shown to exhibit high reproducibility."

8. It is actually unclear whether and why the authors have used MGB only to define the multi-omic aging genes. It would make more sense to discover aging transcripts in the large RNA-seq cohorts (not only the 1,497 transcripts from Peters et al. 2015), discover aging CpGs and genes in the large DNam and RNA-seq cohorts and then validate these genes in MGB, whose

sample size is more modest. Related to this, when are “validated aging genes” used in the study?

Response: The MGB dataset was utilized as our primary discovery cohort despite its modest size because it uniquely offers Epic2 array data with exceptional age range coverage and matched multi-omic profiles from the same individuals. This platform provides superior regulatory element coverage and expanded CpG detection critical for comprehensive epigenetic aging analysis. We acknowledge the methodological considerations regarding cohort size and have added a statement to the discussion:

“We acknowledge the MGB dataset's modest sample size as a limitation, though its EPIC2 platform provided superior CpG coverage critical for comprehensive epigenetic aging analysis. Future studies should integrate EPIC2 technology with larger cohorts to validate our findings and potentially identify additional aging biomarkers that may have been missed due to power constraints.”

9. The authors are right when they state that transcriptomic data show “higher noise, and susceptibility to batch effects” (Meyer et al., Aging Cell 2021). In the context of this study, were the RNA-seq data sets corrected for batch effects?

Response: This is an important point and we indeed normalize expression data in Figure 1(c) where we analyse and combine multiple datasets. We have added a statement to the normalization to the legend of Figure 1:

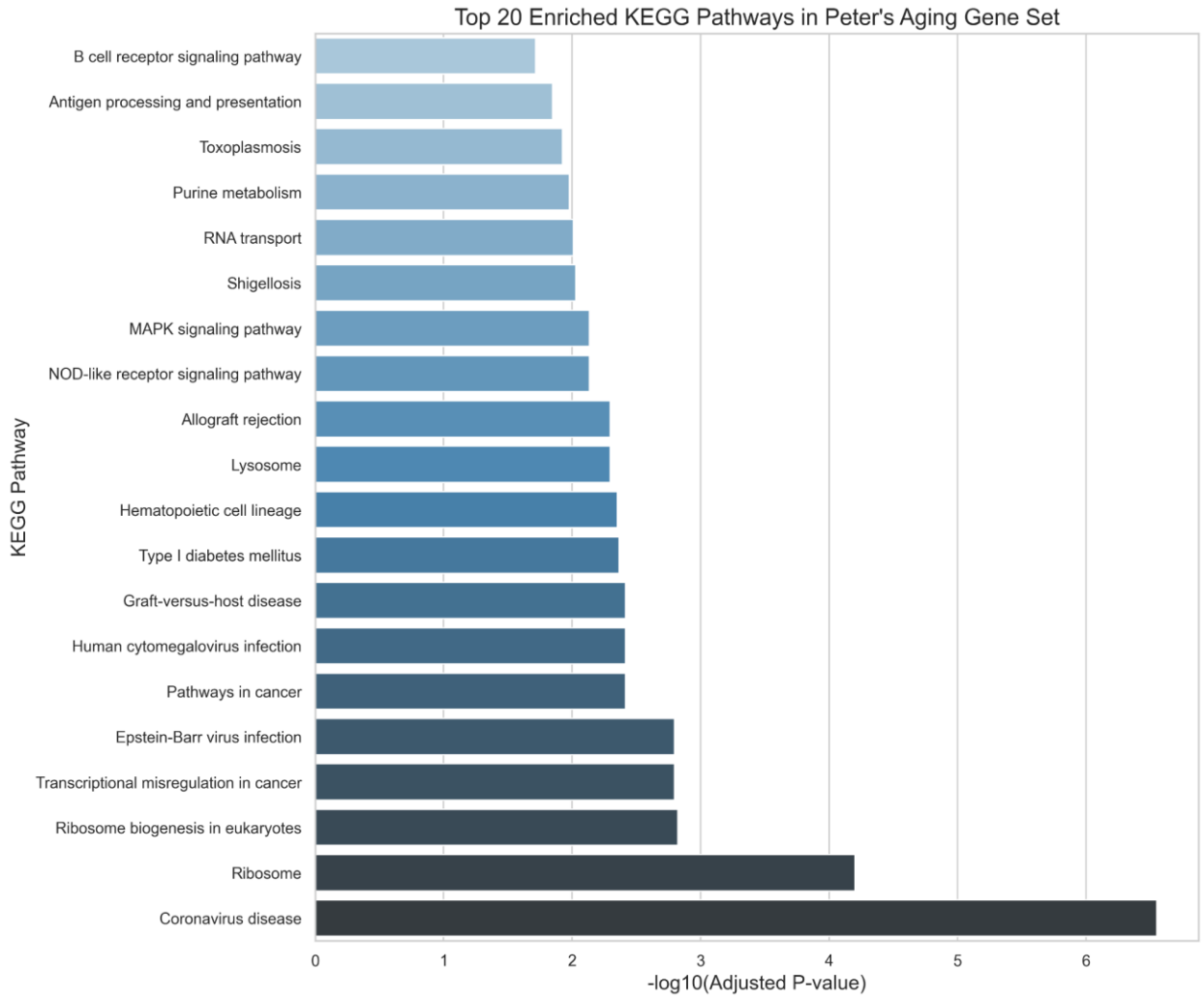
“Gene expression data was normalized using median normalization”

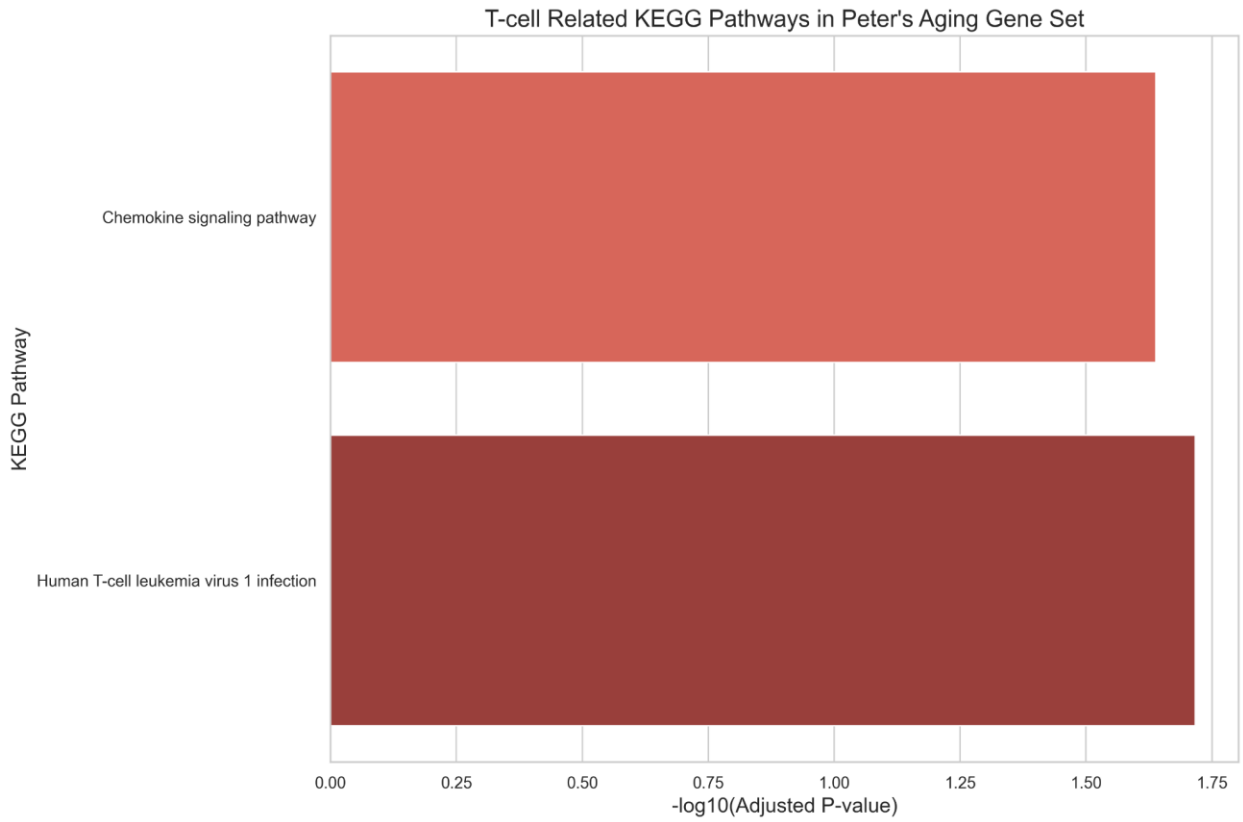
The remaining analyses are done on individual RNA-seq data separately, hence batch effects across datasets become less relevant. We have also adjusted the text in the methods as follows:

“We obtained previous RNA-seq datasets as raw gene expression counts. Count data were normalized using log transformation on (raw counts + 1) and median adjusted to account for batch effects when merging datasets. The majority of analyses was conducted within individual datasets rather than merging multiple datasets.”

10. Are non-“aging genes” with age-associated expression enriched for T cell-specific genes?

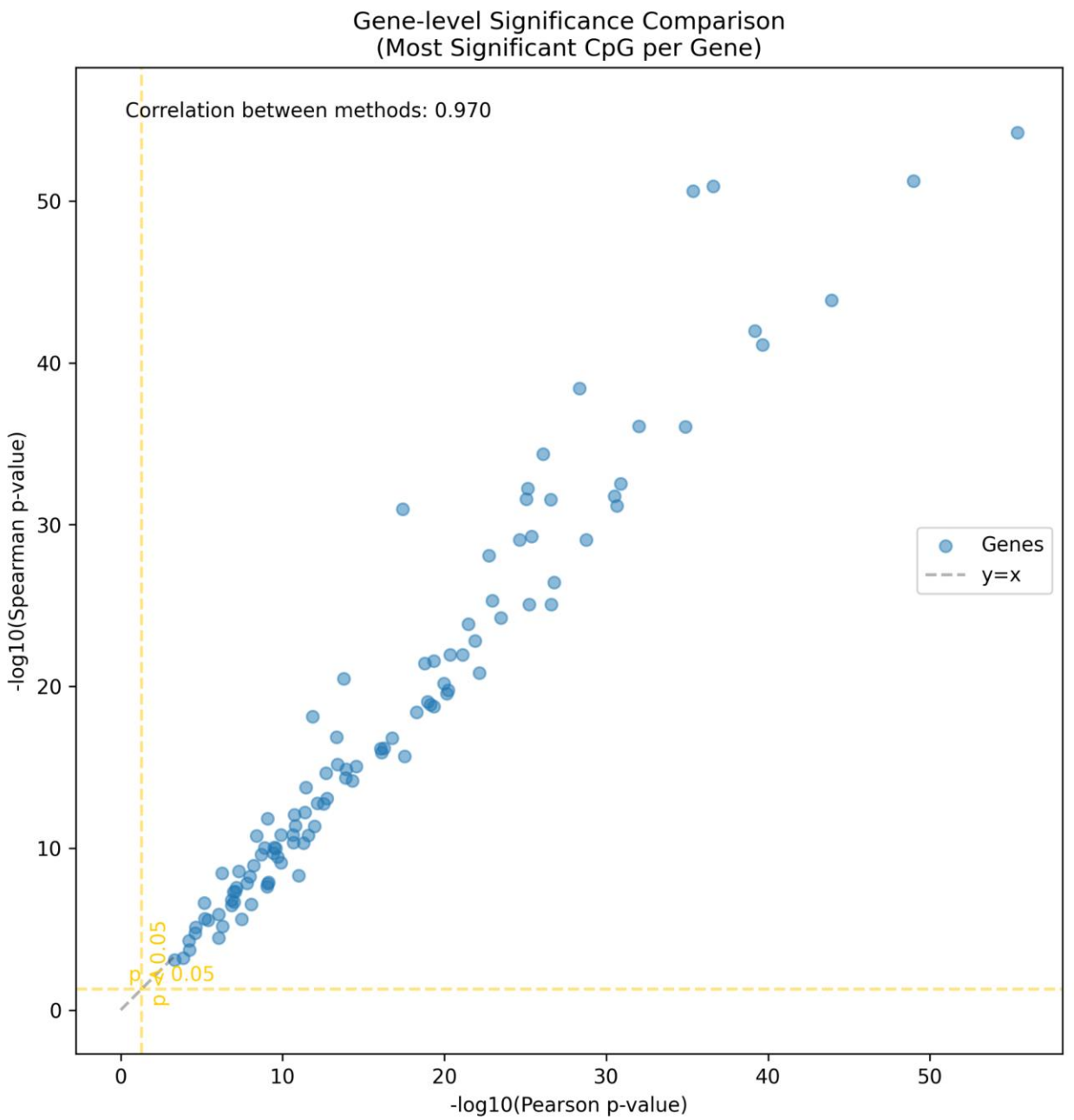
Response: Thank you for this thoughtful question. We performed enrichment analyses for all the genes with age-associated expression identified by Peters' et al and found that they are not in particular enriched for T cell-specific genes as shown below.





11. Outlier values in DNAm and RNA-seq data can result in inflated correlations with age, particularly when the age distribution has long tails (e.g., 500FG, RA). Can the authors confirm their list of 106 aging genes with Spearman correlations?

We thank the reviewer for this methodological point about outlier effects on correlations. Our analysis addressed this concern by calculating both Pearson and Spearman correlations for all CpGs with age in the MGB500 cohort. The high concordance between methods (correlation = 0.970) demonstrates that our findings are robust regardless of correlation method used. All 106 aging genes in our final list show significant correlations using both methods ($p < 0.05$), confirming the reliability of our results independent of potential outlier effects.

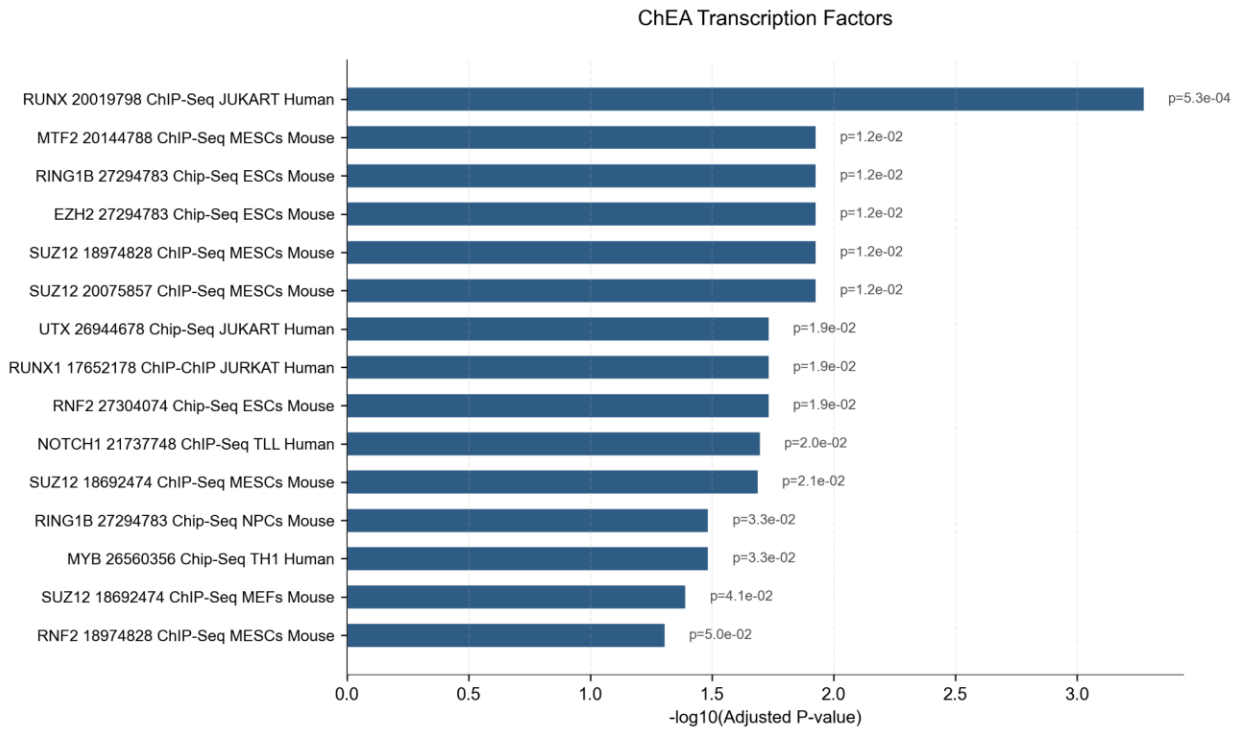


12. It would be very informative to identify further the features that characterize aging genes. Are these genes enriched in binding site for specific TFs?

We thank the reviewer for this excellent suggestion. We have conducted comprehensive transcription factor enrichment analyses. Integration of ChIP-seq data through the ENCODE and ChEA databases revealed significant enrichment for specific transcription factor binding (15 significant TF associations, $p < 0.05$). This analysis identified key upstream regulators

coordinating the expression of our multi-omic aging genes, notably many of which are associated with the PRC2 complex (*EZH2*, *SUZ12*) or immune/haematopoietic cell development (e.g., *RUNX*).

These findings reveal a coordinated regulatory network underlying age-associated epigenetic and transcriptional changes, rather than isolated alterations. This comprehensive regulatory analysis strengthens our understanding of the molecular mechanisms driving aging-associated genomic changes and provides new insights into potential therapeutic targets for age-related conditions.



Minor comments

- I suggest that the authors clarify what is the added value of identifying “aging genes” compared to identifying genes directly associated with age-related diseases.

Response. Thank you for this excellent suggestion. We have included this in the discussion as follows.

> *“We focus on “aging genes” rather than genes associated with specific age-related diseases (e.g., Alzheimer’s disease, cardiovascular disease) to capture upstream, systemic biological processes that drive disease susceptibility across multiple diseases with age and may underlie a shared vulnerability to a spectrum of age-related conditions rather than disease-centric processes.”*

- When the authors state their approach identifies “true aging-associated genes”, can they clarify what they mean?

Response: As we demonstrate, there are many DNA methylation changes without a concordant alteration in gene expression. Here, we identify genes that are jointly altered with both gene expression and DNA methylation, suggesting a more relevant functional role. We have clarified this in the discussion, as follows:

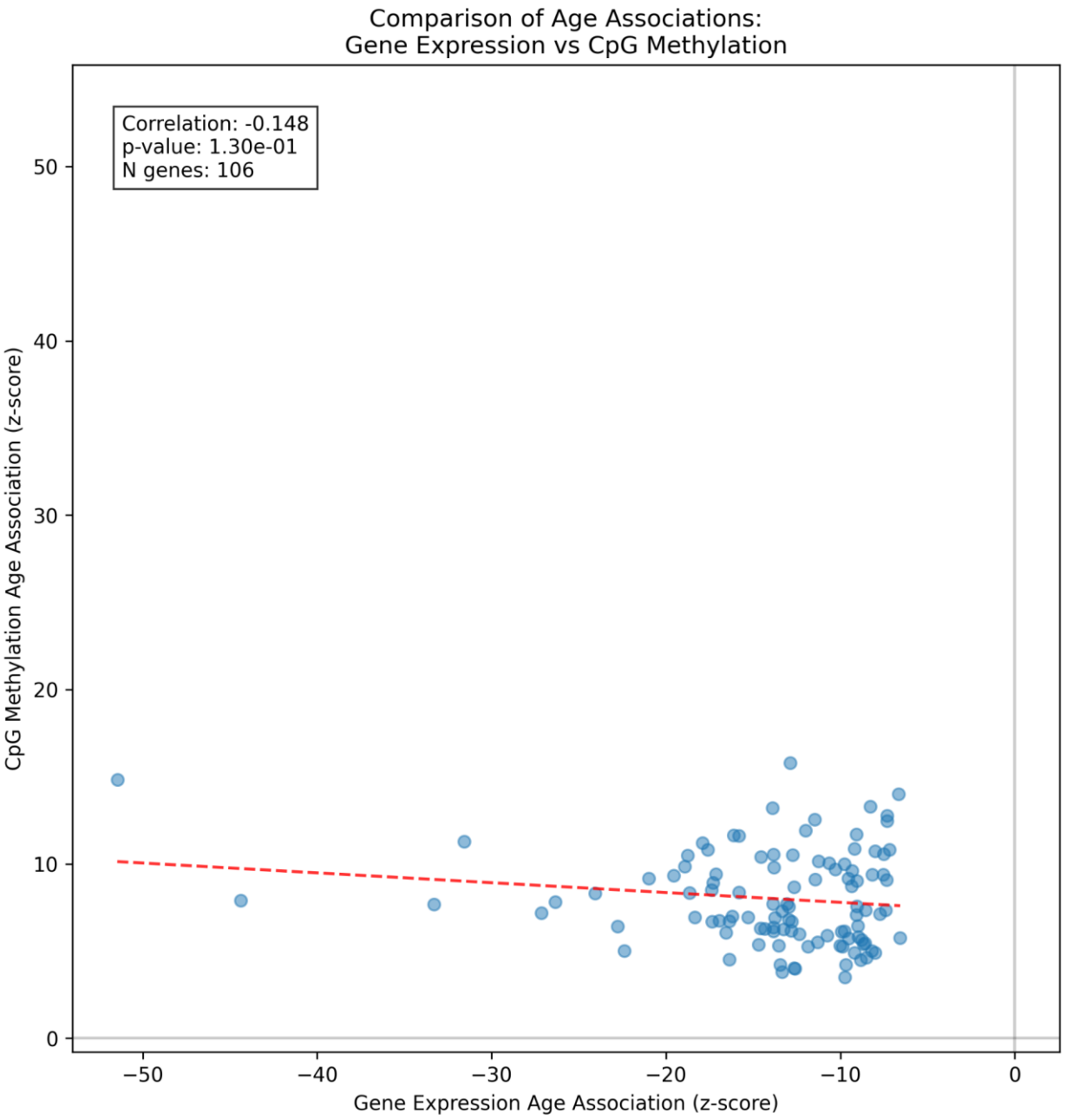
> *“These analyses demonstrate that integration of multiple omics modalities has the ability to refine the identification and validation of true aging-associated genes, where both DNA methylation and gene expression are jointly altered, pointing towards functional shifts”*

- Could the authors speculate further on the causes of the less reproducible effects of age on gene expression across cohorts, compared to DNAm? The CD248 example suggests that age has a non-linear effect on its mRNA levels.

Response: We speculate that the less reliable reproducibility of the association between age and gene expression compared to DNA methylation is partially due to the challenges with measuring transcription levels compared to DNA methylation states. For example many RNA molecules have relatively short life-time while most DNA methylation states are considered to be relatively more stable.

- Fig. 2: It could be informative to plot, for all genes, the correlation with age of gene expression against the correlation for associated DNAm.

Response. Thank you for this suggestion. We have checked the relationship between the age correlation of gene expression and the DNA methylation and overall, there does not appear to be a strong association between these two correlations as shown in the figure below.



- Please italicize gene names in the text and figures.

Response. Thank you for noting this. We have checked all the gene and transcript names and Italicized the names in the text when referring to genes. We have also taken a note of this and will make sure to Italicize all the gene names in the figure vectors that we will produce for the publication.

- In Fig 2c, it would be more informative to show results for PPMI instead of MESA 2010.

Response. In Fig 2c, the results in fact show data for all three cohorts (MESA 2000, MESA 2010, and PPMI).

- Fig 3c, please replace “correlation” by “correlation”

Response. Thank you for noting this. We have corrected the typo in Fig 3c.

- Table 1: Please add information on the tissue in which DNAm and RNA-seq were measured, and the percentage of diseased patients.

Response. Thank you for your suggestion. As mentioned in the text, all the samples are from blood and blood sample types are specified in Table 1, under “cell type”. We have now also added additional details about the participants in each cohort, including the details stated under point #6.

- Methods: Replace “Pearrson” by “Pearson”, and “18,85911 participants” by “18,859 participants”.

Response. Thank you for noting this. We have corrected these typos in the Methods.

- Methods: Please describe in details the quality control filters applied on the new DNAm data.

Response. Thank you for this excellent point. We have added the following details under the Method.

> *“.IDAT files were preprocessed using standard parameters in the R SeSAMe package, version 1.22.1, using the recommended processing for EPICv2 data, with the “QCDPB” parameter (qualityMask, inferInfiniumIChannel, dyeBiasNL, pOOBAH, noob). Samples and CpGs with more than 10 percent missing values were excluded from the generated beta value matrix.”*

Point-To-Point Responses to Reviewers' Comments

Reviewer #1 (Remarks to the Author):

The authors have addressed all my comments and suggestions in full.

We thank the Reviewer #1 for reviewing our manuscript.

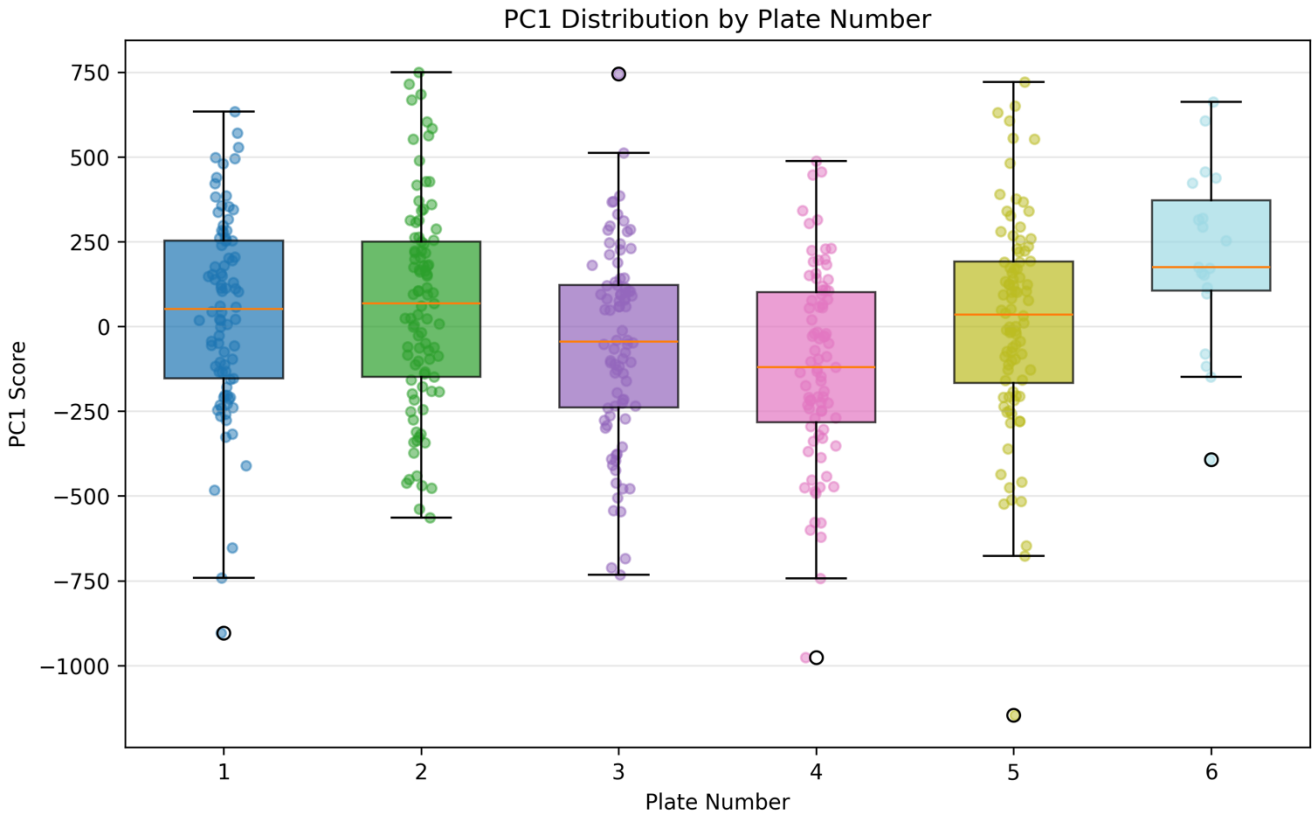
Reviewer #2 (Remarks to the Author):

The authors have satisfactorily addressed most of my comments.

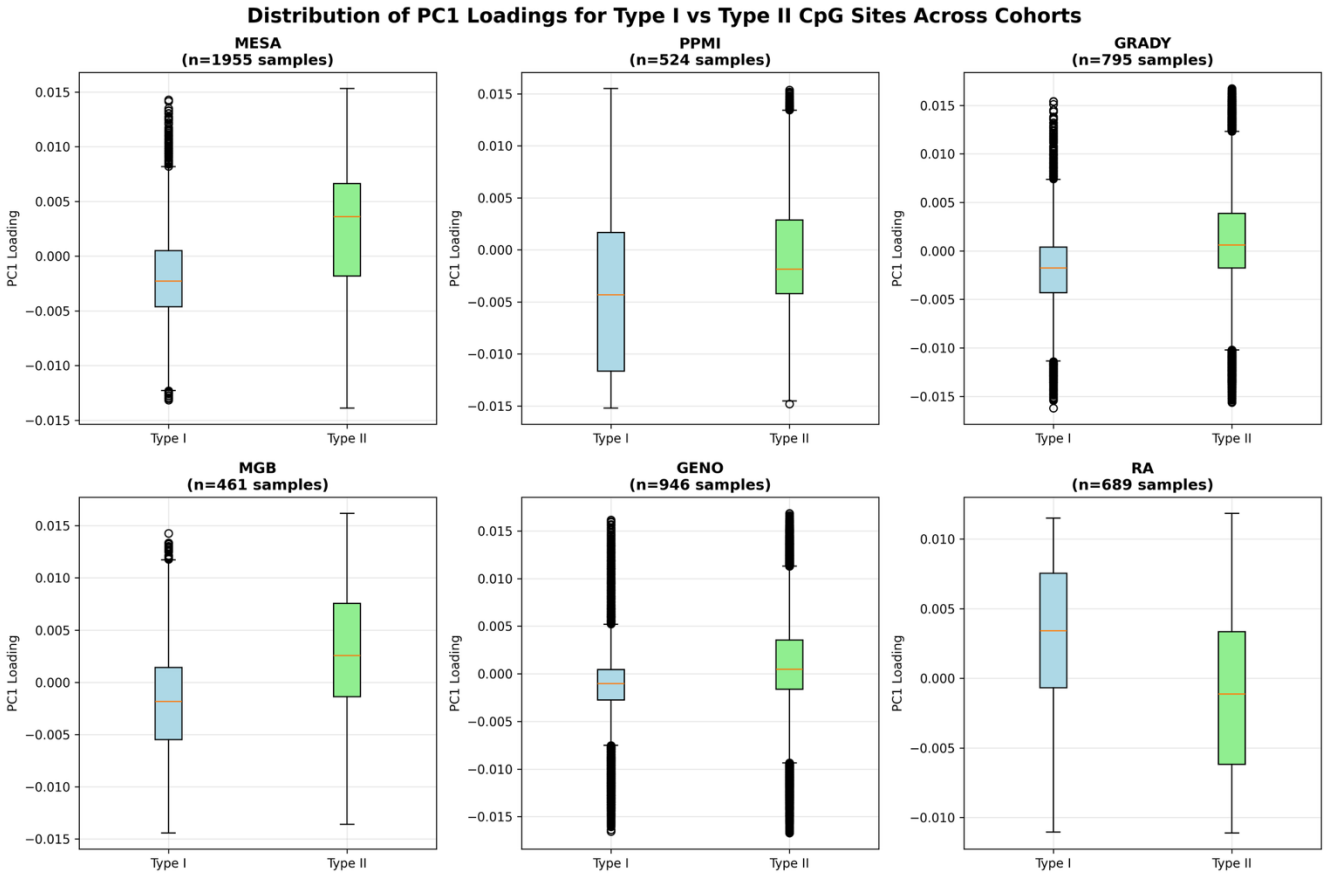
We thank the Reviewer #1 for reviewing our manuscript.

- My only remaining concern is the presence of batch effects within cohorts. Can the authors verify that first PCs of the RNA-seq and DNAm data do not reflect plate effects, beadchip effects or variation in sequencing coverage?

We have performed additional analysis for the quality control in our MGB cohort and observed that the plate numbers are not separated by the PC1.



DNA methylation data from six independent cohorts (MESA, PPMI, Grady, MGB, GenoAge, and RA) were analyzed to investigate the distribution patterns of Type I and Type II CpG sites comparing to PC1. The Illumina Infinium MethylationEPIC array utilizes two distinct probe design types for methylation detection. Type I probes employ two separate probes (one for methylated and one for unmethylated cytosines) that bind to different sequences after bisulfite conversion, providing higher precision but covering fewer sites. Type II probes use a single probe design that interrogates both methylated and unmethylated states at the same locus, offering broader genome coverage but with potentially different technical characteristics. The Illumina Infinium MethylationEPIC array manifest file (v1.0 B5) was used to classify CpG sites based on their probe design type. Of the 865,918 CpG sites on the array, 862,927 sites had type information available, comprising 142,137 Type I probes (16.5%) and 720,790 Type II probes (83.5%). We show that there is no strong bias of Type 1 and Type 2 probes based on the first PC loading.



We have added this part to our Methods section:

Quality control analysis examined the distribution of the plate number as well as Type I and Type II probe loadings on the first principal component across all cohorts to verify absence of systematic technical bias (Supplementary Fig. 13).

All RNA-seq datasets used in this study were already quality-controlled and batch-corrected by the original data producers before we accessed. As we did not have access to the raw sequencing data to independently evaluate technical variation in sequencing coverage, we relied on the preprocessing pipelines implemented by the respective consortiums. TOPMed (MESA) datasets underwent standardized laboratory methods with centralized processing using single pipelines to minimize batch, and PPMI RNA sequencing data were uniformly processed with standardized pipelines from RNA isolation through gene detection to control technical. The original publications for these datasets report comprehensive batch effect correction procedures, including empirical Bayes methods and variance stabilization approaches appropriate for RNA-seq count data.

Our analysis approach of conducting the majority of analyses within individual cohorts rather than merging datasets further minimizes cross-cohort technical artifacts, with robust cross-cohort validation serving as an additional control against batch-driven findings.

- L. 123: Please italicize gene names.

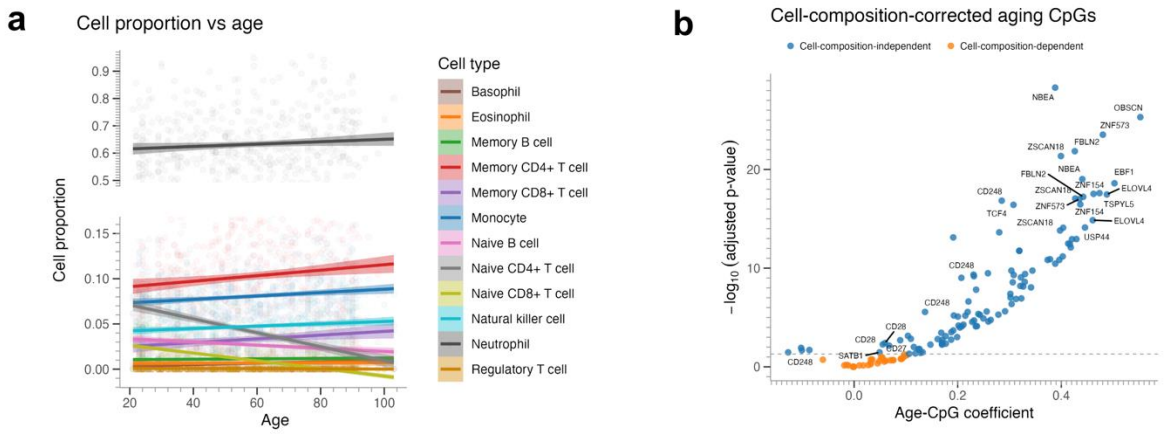
We have italicized gene names.

- Should the authors remove sentence in L. 130-131?

We have removed it.

- Resolution of Suppl. Fig. 12 should be improved. Could the authors clarify which CpGs were selected for analysis shown in panel b?

We have updated Supplementary Figure 12 to improve the resolution. The graph includes all the CpGs we identified in our analysis (around the TSS of the 104 aging genes). We've also added this description to the legend.



Supplementary Fig. 12. Analysis of cell type composition and its impact on age-CpG associations. a. Relationship between chronological age and estimated cell proportions for 12 different immune cell types. b. Scatter plot showing the association between age and CpG methylation after correction for 12 cell types. Significant associations remained after adjustment, indicating that age-related changes in these CpGs are largely independent of changes in cell composition. All the CpG sites around the TSS of the 104 aging genes we previously identified are included in this analysis.

- The Discussion is reported twice.

We have removed the extra copy of the discussion.