GENE NETWORK ANALYSIS REVEALS FUNCTIONAL DRIVERS OF AGE-ASSOCIATED DISORDERS

Jacob Bridge (bridg245), Aaron Castle (castl145), Annabelle Coler (coler018)
Isabel Constable (const121), Kyla Hagen (hagen772), Evan Pochtar (pocht004)

University of Minnesota Twin Cities

May 2024

Abstract

Due to a progressive disruption of cellular homeostasis and communication, chronic disease risk increases dramatically with age. As a result, according to the "Geroscience Hypothesis," therapies that directly target drivers of aging could dramatically improve geriatric health outcomes. Although many age-associated pathways have been described, there is significant disagreement regarding their relative contributions to distinct tissue and disease phenotypes, complicating the drug development process. We hypothesize that drivers of aging can be inferred through gene network analysis, as the upstream regulators responsible for coordinated patterns of age-dependent expression. By applying regression models to RNA-seq data from over 3,000 human donors, encompassing myriad ages, tissues, and disease states, we isolate tissue-independent expression changes associated with healthy aging. We then identify characteristic variations associated with 16 geriatric diseases, and construct an interactive network of coordinately regulated gene sets. Critical nodes within the network display substantial functional agreement with known hallmarks of aging, reinforcing the utility of our approach. Annotation of network outputs also allows for a granular assessment of putative drivers, with the potential to streamline identification of robust therapeutic targets.

Introduction

Aging is characterized by a breakdown in the macromolecular structures required for organismal function, leading to the dysregulation of metabolic pathways and disruption of cell and tissue homeostasis.[1] As a result, old age is accompanied by an increase in morbidity, most notably for chronic conditions such as cancer, heart disease, and neurodegenerative disorders.[2] While many of these diseases have robust genetic and environmental associations, such as smoking with lung cancer, the strongest predictive factor for nearly all of them is age itself.[3] This serves as the thrust behind the "Geroscience Hypothesis," which posits that because aging physiology plays a major role in most chronic diseases, therapies that directly target drivers of aging could dramatically enhance quality of life.[4] Translational applications of this paradigm, including the use of senolytic drugs to combat senescent cell accumulation, are currently under development.[5]

Despite recent advances, the success of these therapies will ultimately rest on a robust understanding of the factors that drive aging phenotypes. Without an accurate characterization of the roles played by each biomolecular component, it would be nearly impossible to distinguish relevant drug targets from the sea of potential alternatives. The aging field has already zeroed in on many putative drivers over the course of its history, ranging from telomere length[6] and oxidative damage[7] to metabolic dysfunction[8] and somatic cell mutations.[9] While each element has its own backers and detractors, none has, on its own, produced a model that can fully describe human aging. Over the last decade, efforts to consolidate these functional drivers have produced consensus hallmarks of aging,[10] distilling them into nine–and more recently twelve[11]–categories. However, the relative importance of targetable pathways within each hallmark, along with their contributions to distinct tissues and disease states, remains elusive.

Here we report that functional drivers of aging can be inferred through a network analysis of gene expression data. Current evidence suggests that aging phenotypes stem from a combination of molecular disruptions that span the breadth of cellular structure and function, rather than alterations of a few individual proteins. Thus, we hypothesized that targetable pathways associated with aging phenotypes would emerge as conserved upstream regulators of

coordinately regulated, age-dependent gene sets. Using RNA-seq data from over 3,000 human donors, representing a broad array of ages, tissues, and geriatric diseases, we identified consensus expression patterns associated with healthy and disordered aging. We then assembled these components into an interactive network, enabling the resolution of therapeutic targets associated with hallmarks of aging.

Methods

Data Subsetting and Normalization

Raw RNA-seq counts and associated metadata were collected from a meta-analysis by Shokhirev and Johnson (2021).[12] Summary information, including the age distributions associated with each tissue and disease state, was computed from the metadata. Count data was first normalized by counts per million (CPM) to control for variations in sequencing depth and library size across samples, followed by a transcripts per million (TPM) adjustment to control for differences in gene length. Variance stabilization and data normalization was accomplished using a log2 transformation. Batch correction was performed using ComBat analysis. All data manipulations were performed in R.

Linear Regression

Following normalization, tissue-independent changes in age-associated gene expression were calculated using the multiple linear regression model

$$log_2(E) = Age + Tissue = \beta_0^A + \beta_1^A A + \beta_1^T U_1 + ... + \beta_n^T U_n$$

where E is the expression level of a given gene, A is a patient's age, and $U_1 ... U_n$ are dummy variables representing tissue of origin. Input data was subsetted to include only healthy patients. To assess coordinated age-associated changes in the expression of discrete functional pathways, genes were assembled into curated gene sets found within collection C2 of the Molecular Signatures Database.[13] Linear aging coefficients for gene sets were generated by computing average TPM values for each healthy patient, then plugging the resulting mean expression vector into the above model. Age-associated slope coefficients (β_1^A) and their corresponding p-values values were used to identify age-dependent genes and gene sets. Bonferroni corrections were used to ensure expression changes were significant, while thresholds of $|log_2(E)| > 0.05$ and $|log_2(E)| > 0.02$ were imposed for genes and gene sets respectively as benchmarks for biological relevance.

Disease-Associated Differential Expression Analysis

To assess whether geriatric diseases produce characteristic alterations in age-dependent gene expression, expression levels of age-associated genes and gene sets were compared between healthy and diseased samples. Diseases with an insufficient sample size ($n \leq 2$) were eliminated from consideration, leaving 16 eligible conditions. To prevent sample age from confounding disease-specific effects, healthy tissue data was subsetted to exhibit an identical median age to disease data. For each pairwise expression comparison between healthy and diseased samples, a two-sided t-test was conducted. By-disease differences in gene and gene set expression level, along with their corresponding p-values, were then calculated and stored in matrices. The data were exported as a CSV file and used for the network analysis. A summary analysis of consensus differentially-expressed gene sets across all diseased tissues, irrespective of the significance of their age-associated slope coefficients, was also performed.

Assessment of Co-Regulatory Interactions

To identify clusters with similar age-dependent expression patterns, Pearson correlation coefficients (PCCs) were calculated for each pairwise combination of age-associated gene sets across all healthy patients. Average expression values for each gene set were used as inputs. To more effectively compare the distribution of gene expression values within each set, Jensen-Shannon divergence (JSD)[14] was computed for each combination of samples and gene sets. P-values from a linear regression model to assess whether JSD values changed significantly with time were also recorded.

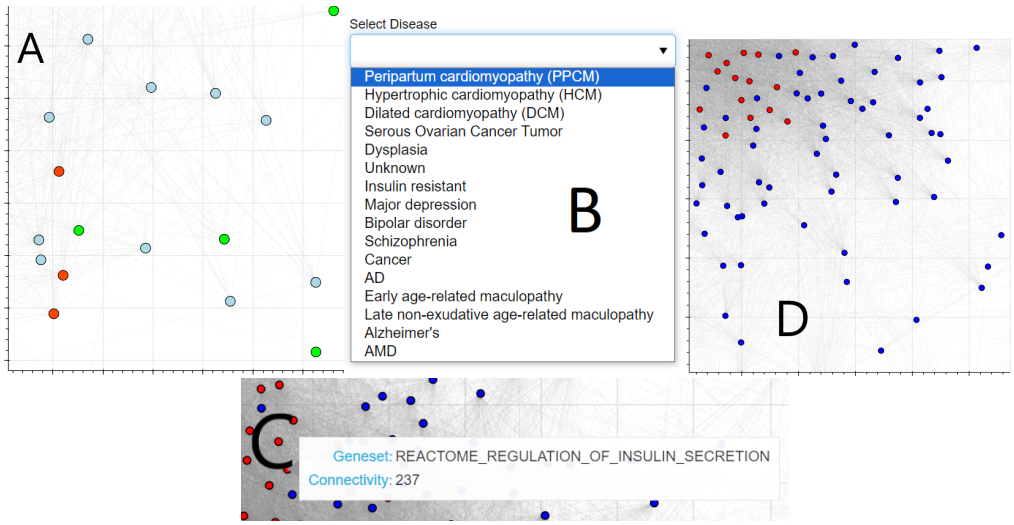

Figure 1: a) A cut of the network graph showing regulation for early age-related maculopathy. A node being lime colored represents down regulation when the disease is present, a red colored node represents up regulation, and a light blue colored node represents no notable difference. b) The selector and options for which diseases can be chosen. The network graph is automatically updated on click. c) Example of the hover functionality that presents the geneset name and connectivity when hovering over a node on the network. d) A snippet of the network, red represents the top 100 most connected nodes, blue is every other node.

Network Creation

A holistic network of age-associated gene sets was created in python, using the networkx spring layout for display. Gene sets were represented as nodes, with edges reflecting |PCC| values above 0.8. The 100 most interconnected gene sets were highlighted in red, while all others appeared blue (Figure 1d). Using Bokeh, the network was then transferred into HTML, enabling several interactive features. "HoverTool" mechanisms were implemented to display gene set name and connectivity when hovering over an individual node (Figure 1c). Disease association data was also incorporated, allowing users to view altered expression patterns corresponding to different disorders. A simple javascript program was written to detect a change in the Bokeh "Select" tool (Figure 1b) and display the information for the chosen disease. Node color was coded to change based on whether each gene set was upregulated (red), downregulated (green), or expressed equally (light blue) in that condition relative to healthy tissue (Figure 1a). A second network, with edges reflecting JSD values below 0.2 and p-values above the Bonferroni significance threshold, was also generated.

Annotation of Functional Drivers

Hallmark of aging (HOA) classifications for over 6,000 genes were accessed from paper by Holzscheck et al. (2020).[15] To assess the robustness of our regression model, age-associated gene sets were assigned HOA components based on the fraction of corresponding genes annotated within each hallmark. HOA components were then compared to those derived from a random classifier. Relative enrichment of HOAs within the top 100 gene sets was also computed, with the HOA component of each critical node being weighted by its connectivity. To assess putative drivers of aging in a more granular fashion, the full descriptions of each age-associated gene set were scraped from MSigDB using the rvest package. The long descriptions were searched for a list of various keyword annotations associated with HOAs, and the number of gene set descriptions found containing each keyword were recorded.

Results

Summary of RNA-seq Data

The age distributions of each disease condition, presented by decreasing number of samples, showed a wide range of ages of healthy samples, spanning nearly 100 years (Figure 2, Supplementary Table 1). Each of the diseases had a relatively small distribution of ages, with most of their IQR having less than a 20 year spread.

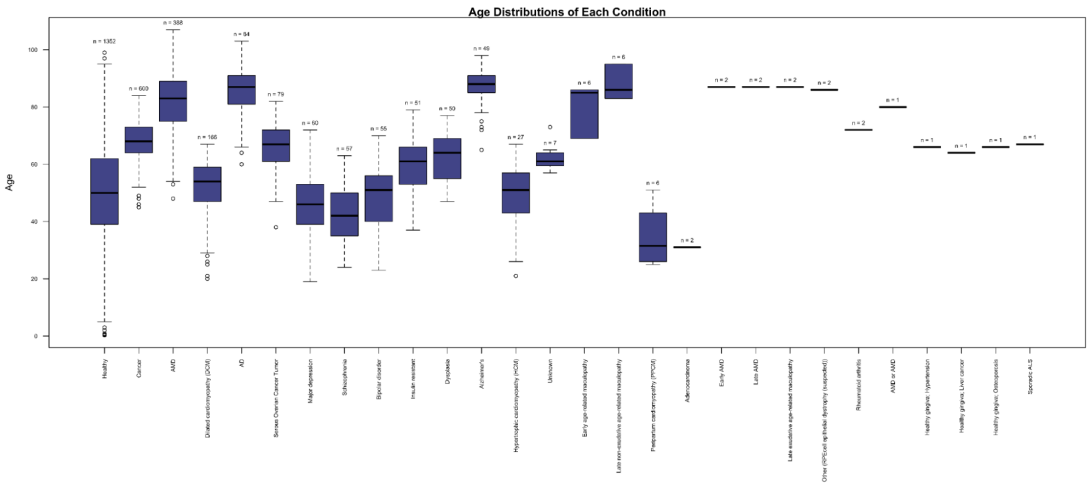

Figure 2: Age distributions of each condition. Boxes show the interquartile range (IQR) of sample ages for each condition. The black lines within each box denote the median age. The dotted lines mark 1.5 times lower or higher than the IQR. Values beyond those bounds are considered outliers and are marked with open circles. If no values exceed the 1.5 times IQR, the whiskers denote the minimum and maximum values. Above each box is the number of samples, n, of each condition.

The age distributions of each tissue, also organized by decreasing number of samples, is shown in Figure 3 and Supplementary Table 2. Non-small cell lung cancer (NSCLC) made up the largest number of samples, with n = 600. Tissues with larger sample sizes had a relatively small distribution of ages, so we determined it was not useful to examine the effect of age on individual tissues, but to instead focus on tissue-independent analysis.

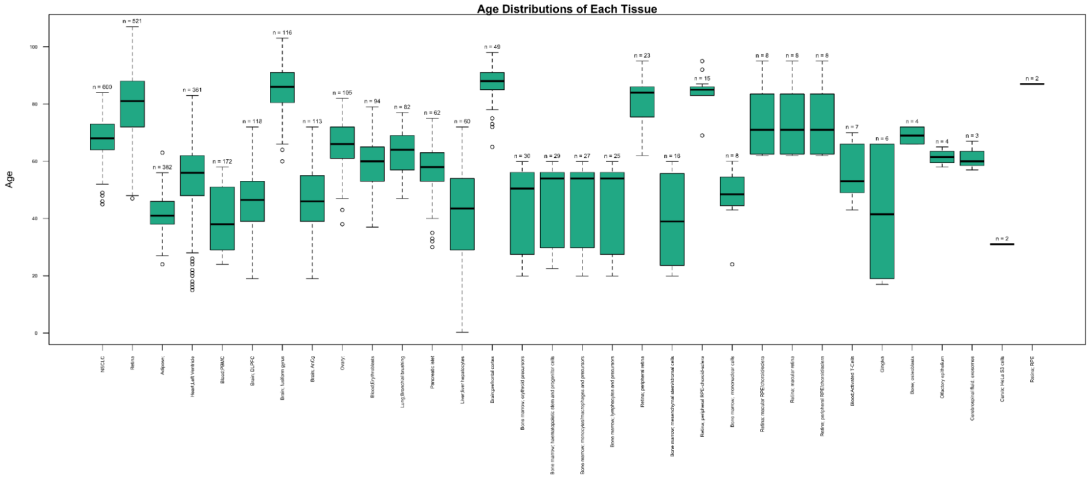

Figure 3: Age distributions of each tissue. Boxes show the interquartile range (IQR) of sample ages for each tissue. The black lines within each box denote the median age. The dotted lines mark 1.5 times lower or higher than the IQR. Values beyond those bounds are considered outliers and are marked with open circles. If no values exceed the 1.5 times IQR, the whiskers denote the minimum and maximum values. Above each box is the number of samples, n, of each tissue.

Identifying Age-Associated Genes and Gene Sets

After subsetting the data to include only healthy tissue samples (n = 1355), linear regression analysis produced 1149 significantly differentially expressed genes with a biologically relevant change in age-dependent expression ($|log_2(E)| > 0.05$; Fig. 4A). Genes were successfully assembled into MSigDB gene sets (n = 6366), yielding 566 age-associated gene sets. The biological relevance threshold for gene sets was determined to be lower than for genes ($|log_2(E)| > 0.02$; Fig. 4B), due to the stabilizing effect of much larger sample sizes. The vast majority of differentially expressed genes and gene sets were upregulated with age.

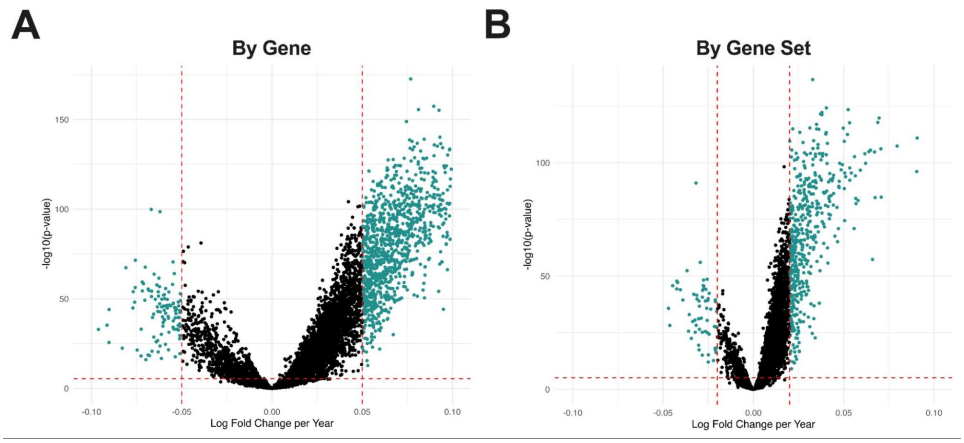

Figure 4: Summary of tissue-independent linear regression output. (A) Log2-fold change in gene expression by year. The horizontal red line indicates the Bonferroni correction threshold, while the vertical red lines at -0.05 and 0.05–equating to a 2-fold change in gene expression every 20 years–represent the threshold for biological relevance. (B) Log2-fold change in gene set expression by year. The thresholds for biological relevance were adjusted to ± 0.02.

Age-Associated Expression Patterns Vary by Health Status

The regression-derived tissue-independent pathway descriptions were examined on MSigDB. The curated gene sets collection (C2) contains the subcollections canonical pathways (CP) and chemical and genetic perturbations (CGP) subcollections. Canonical pathways were found to dominate both the top up-regulated and top down-regulated gene sets (Supplementary Table 3, Supplementary Table 4). The fourth most up-regulated pathway, and the only top ten pathway from the CGP subcollection, was involved with regulation of gene expression. The remainder of the top ten up-regulated tissue-independent pathways were related to vision and neurotransmitter receptors. The top down-regulated pathway, also the only pathway from the CGP subcollection, played a role in male adult germ cell tumors. The remaining nine of the top down-regulated pathways were all related to the function of the immune system.

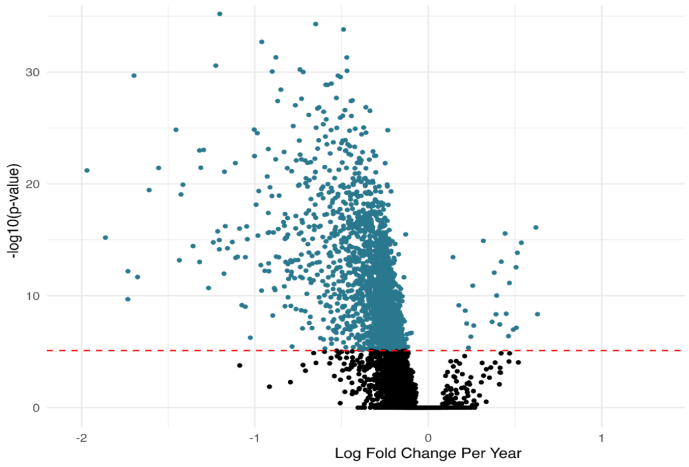

Figure 5: Age-associated gene expression changes differ across disease states. Differential gene set expression in diseased tissues relative to healthy ones. The horizontal red line indicates the Bonferroni correction threshold. Below the line indicates samples that are not significant.

Most of the data points cluster around zero on the x-axis, suggesting no significant change in expression for these genes (Figure 5a). There is a dense cluster of points along the zero vertical axis below the significance threshold on the y-axis, indicating non-significant p-values for these gene changes (Figure 5b). Points that appear in the upper regions of the plot (above the red dashed line) are the genes considered significantly differentially expressed, with some showing substantial upregulation or downregulation as indicated by their distance from zero on the x-axis (Figure 5b).

AMD shows the most upregulated genes and gene sets out of all the diseases, where the count of upregulated gene sets significantly exceeds those that are downregulated (Figure 6b). Bipolar, Major Depression, Hypertrophic cardiomyopathy (HCM) and Peripartum cardiomyopathy (PPCM) only exhibit downregulation for both genes and gene

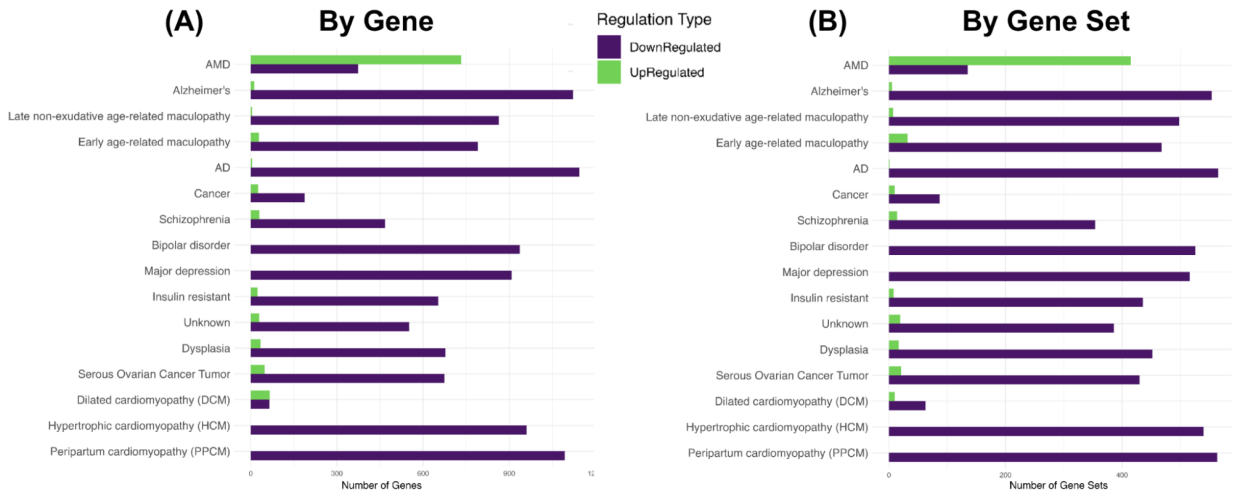

Figure 6: Age-Associated Expression Patterns Differ by Disease: a) Number of up-regulared (green) and down-regulated (purple) age-associated genes by disease type. b) Number of up-regulared (green) and down-regulated (purple) age-associated gene sets by disease type.

sets (Figure 6). Alzheimer's, late non-exudative age-related maculopathy, AD, cancer, schizophrenia, insulin resistant, dysplasia, serous ovarian cancer tumor, and dilated cardiomyopathy show a substantial number of downregulated genes/gene sets with slight upregulation (Figure 6).

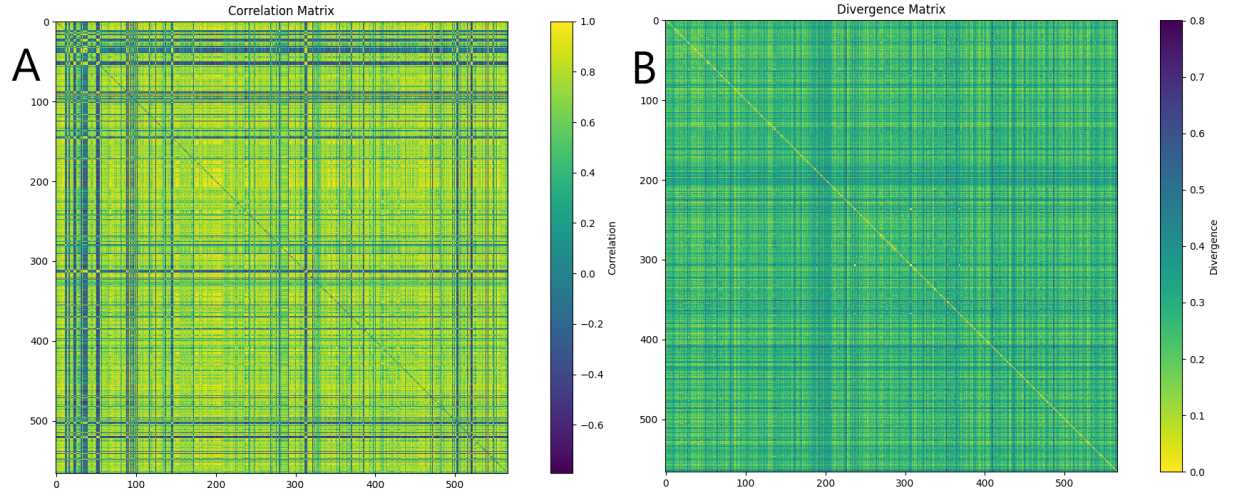

Figure 7: Correlation matrices for age-associated gene sets. (A) Pearson correlation coefficients (PCCs) for pairwise comparisons of average gene set expression across healthy tissue samples. (B) Average Jensen-Shannon Divergence (JSD) values corresponding to the pairwise gene set comparisons in (A).

Assessing Gene Set Co-Regulation

After calculating PCC values for each pairwise combination of age-associated gene sets, we observed generally high correlation values throughout the resulting matrix (Fig. 7A). The vast majority of correlation values were positive, suggesting an abundance of positive co-regulatory interactions amongst age-dependent pathways. Mean JSD values were concentrated within the low to intermediate (0.2-0.4) range (Fig. 7B), reinforcing the generally high agreement in expression values between age-associated gene sets. Overall, variation across the PCC matrix, although relatively low, still far exceeded that of the JSD matrix (0.0385 vs. 0.00718), even after compressing the PCC scale by a factor of two.

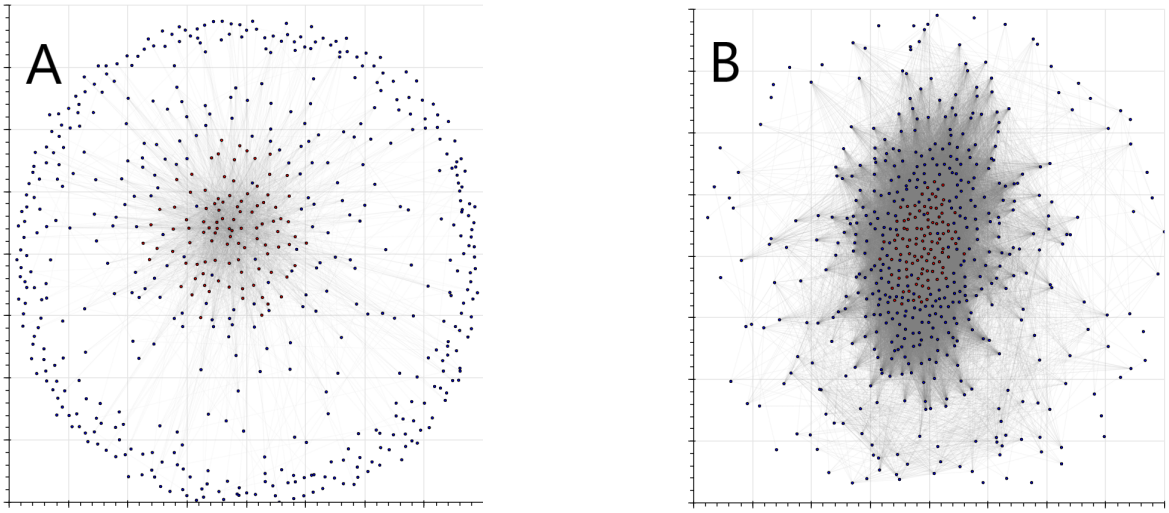

Figure 8: a) JSD Network, edges defined as any geneset having a JSD mean lower than 0.2 and a p value higher than 3.127e-7. Edges will then represent gene set associations that remain consistent with age and exhibit low overall divergence. b) This network is created using Pearson Correlation data from genesets differentially expressed in aging. An edge is defined by anything with a correlation within $-0.8 \geq x \geq 0.8$.

Developing a Network of Age-Associated Gene Sets

After finding gene sets differentially expressed in aging, each gene set was placed into a network for further analysis. Initially, the team tried a JSD approach. This consisted of defining an edge as any gene set having a JSD mean lower than 0.2 and a p-value higher than 3.127e-7. This produced a network graph that had many outliers with zero, or close to zero, connections to any other gene set on the graph (Figure 8a). Edges then represent gene set associations that remain consistent with age and exhibit low overall divergence. Any adjustments to the mean value to increase the amount of connections resulted in further clustering in the middle of the network, without much effect of reducing the outlier population. Although the gene sets the graph clustered around were promising and seemed relevant to the hypothesis, the team decided it was too unreliable of a data source. Instead, a Pearson correlation approach was implemented. In this case, an edge between two nodes is defined as anything with a correlation greater than .8, representing a positive edge, or -.8, representing a negative edge. This produced a much more clustered graph, where almost every node was connected to another (Figure 8b). This was the result the team expected, as the way the gene sets were derived from the initial data was by finding differentially expressed gene sets. This resulted in a set of gene sets that all increased or decreased at the same time, increasing the overall Pearson Correlation throughout the data. This indicates that the Pearson correlation gene set network is producing the intended results. The network allowed for inspection of connected nodes associated to different hallmarks of aging, as well as utilizing the HTML functionality to allow for further inspection of the network by using disease data to find connected nodes down or upregulated in certain diseases.

Characterizing Network Outputs by Hallmarks of Aging

To assess the performance of our linear regression model, we cross-referenced the members of each age-associated gene set with a collection of 6642 genes linked to HOAs. Compared with a random classifier (Fig. 9A), age-associated gene sets were significantly enriched for HOA annotations ($p = 0.00153$; Fig. 9B), particularly deregulated nutrient sensing and stem cell exhaustion. We also assessed the relative representation of each HOA within critical nodes in the network, weighting the HOA fractions for each top 100 node by their corresponding connectivity (Fig. 9C-9D). Altered intercellular communication accounted for the nearly 40% of edge-weighted HOA annotations within the critical node subnetwork, while cellular senescence, epigenetic alterations, and telomere attrition accounted for the smallest fraction of annotations.

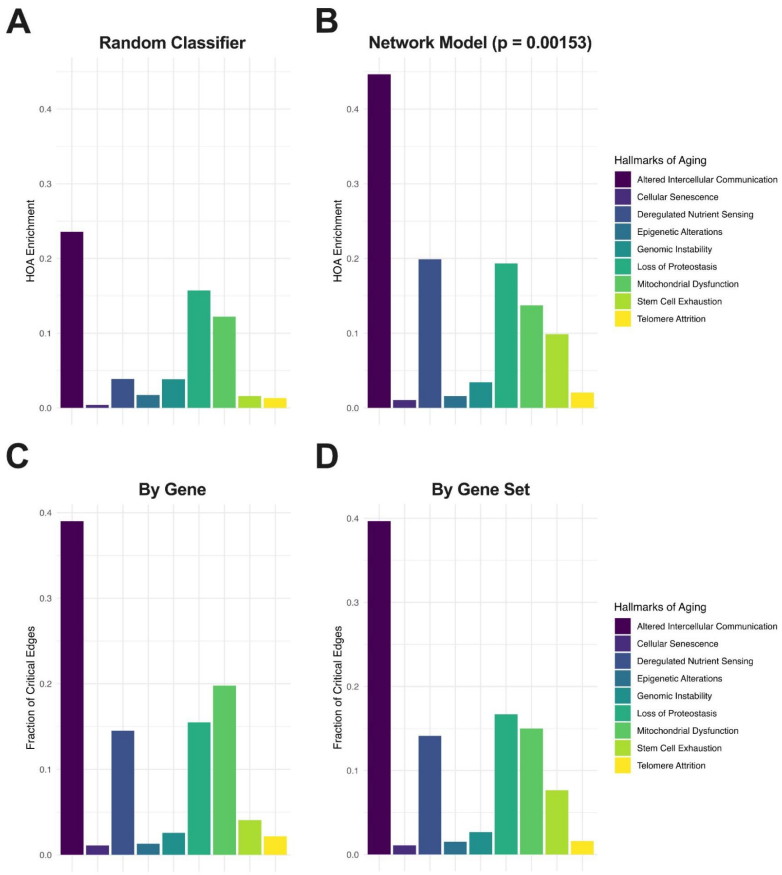

Figure 9: Age-associated linear regression and network outputs exhibit substantial agreement with hallmarks of aging (HOA). (A) HOA annotation using a random classifier. Expected values were computed based on the fraction of literature-characterized HOA genes relative to total genes in the raw data file. (B) Enrichment of HOA within the age-associated gene sets used for network construction. (C) Edge-weighted representation of hallmarks of aging within the top 100 most interconnected gene sets, normalized by the number of genes in each set. (D) Non-normalized (by-set) data corresponding to (C).

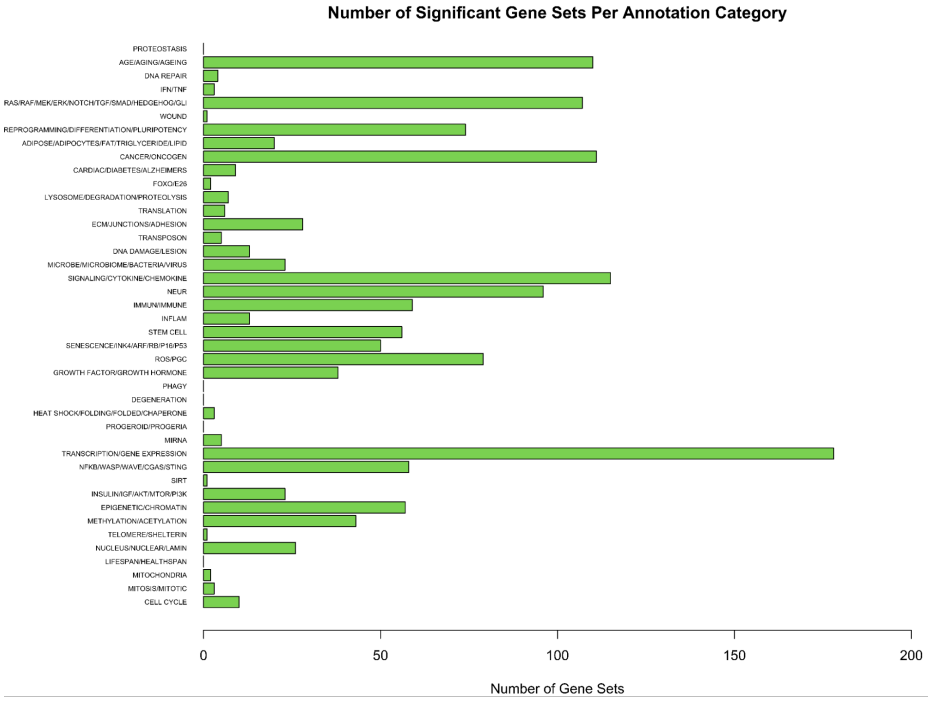

Figure 10: Significantly differentially expressed gene sets belonging to keyword annotation categories. Keywords shown on the y-axis exactly as used to search gene set descriptions. Partial words were used to hit any derivatives. Multiple words indicated relation to the same target category.

Assessing Putative Drivers of Aging

We searched the MSigDB descriptions of each significantly differentially expressed gene set from the regression-derived network and pulled relevant gene sets compared to a list of known age-associated keyword terms and pathways. Terms that were enriched in these gene sets included: aging, the Ras/Raf/MEK/ERK signaling pathway, cancer, stem cells and differentiation, chemical signaling, senescence, reactive oxygen species, and gene expression (Figure 10).

Discussion

Defining the functional drivers of aging will be critical to development of future geoscientific therapeutics. Here we demonstrated a network-based approach to identify conserved age-associated pathways, allowing for streamlined identification of robust drug targets. We observed that unhealthy patients in our dataset were nearly 20 years older on average than healthy ones, reinforcing the well-characterized link between old age and chronic diseases.[2] We also determined that by-tissue analysis was not useful based on the age distributions of each tissue having small ranges and/or small sample sizes. Instead, we focused on the development of a tissue-independent regression model that could identify consensus drivers of aging across a broad array of tissues.

The most highly up-regulated gene sets with age were nearly all associated with ocular homeostasis, while the most down-regulated gene sets were consistently associated with immune function. While the latter aligns with a well-characterized reduction in immune function with age,[1,3,4] explanations for the former are less obvious. Macular degeneration was also the only chronic disease to exhibit predominantly upregulation relative to healthy tissue, which could reflect a distinct eye-specific aging trend. Regardless, more research will be needed to validate this finding, and ensure it is not simply an artifact within the data.

When generating correlation coefficients, we observed that the overall variance within the JSD matrix was much lower than that of the PCC matrix. This appeared to significantly hinder the utility of its outputs, as edge thresholds such as 0.2 produced excessively sparse networks, while higher thresholds produced networks that were too tightly interconnected to resolve useful functional data. While JSD data should theoretically generate a more robust assessment of gene set correlation, it appears to be broadly impractical within our current pipeline. It is possible that the use of a machine learning approach to assess variance within and between pairwise combinations of gene sets could produce a correlation matrix that better reflects the intricacies of genetic data.

Nevertheless, we identified a significant enrichment of HOA annotations within our network relative to a random classifier. In addition to validating the robustness of our approach, this also provided a window into the characteristics of pathways associated with each hallmark. While altered intercellular communication encompasses a highly integrated set of functional drivers associated with myriad cellular processes, telomere maintenance, and its corresponding attrition, is regulated by a small, insulated complex of shelterin proteins.[10] This may explain the observed differences in critical edge representation, despite both HOAs being enriched to a similar extent–roughly two-fold–relative to a random classifier.

Lastly, as a proof of concept, we cross-referenced our network with a broad array of pathways and cellular components implicated in aging phenotypes, with dramatic enrichment of certain terms relative to others. While this may help indicate relative biological importance, our pipeline will have to be further assessed and refined to confirm the functional relevance of these outputs.

Overall, we have demonstrated that gene expression changes with age across a broad array of human tissues. We have also shown that age-associated disorders exhibit distinct gene expression profiles. Network-derived gene sets overlap substantially with known hallmarks of aging and align with other putative drivers. In the future, we hope to expand our hallmarks of aging annotations to include individual disease states. This could aid in the development of drug treatments to target particular chronic diseases, even in the absence of a broader geroscientific approach. We would also like to use deep learning to assess co-regulation in a more nuanced way than the use of Pearson correlation coefficient. Finally, we hope to develop an improved model to describe complex changes in age-associated gene expression, resulting in a network capable of more robust functional predictions.

References

[1] DiLoreto, R., and Murphy, C. T. (2015). The cell biology of aging. Mol Biol Cell 26(25), 4524–4531. 10.1091/mbc.E14-06-1084.

[2] McCune, S., and Promislow, D. (2021). Healthy, active aging for people and dogs. Front Vet Sci 8, 655191. 10.3389/fvets.2021.655191.

[3] Niccoli, T., and Partridge, L. (2012). Ageing as a risk factor for disease. Curr Biol 22(17), R741–R752. 10.1016/j.cub.2012.07.024.

[4] Sierra, F., Caspi, A., Fortinsky, R. H., Haynes, L, Lithgow, G. J., Moffitt, T. E., Olshansky, S. J., Perry, D., Verdin, E., and Kuchel, G. A. (2021). Moving geroscience from the bench to clinical care and health policy. J Am Geriatr Soc 69(9), 2455–2463. 10.1111/jgs.17301.

[5] Power, H., Valtchev, P., Dehghani, F., Schindeler, A. (2023). Strategies for senolytic drug discovery. Aging Cell 22(10), e13948. 10.1111/acel.13948.

[6] Shammas, M. A. (2012). Telomeres, lifestyle, cancer, and aging. Curr Opin Clin Nutr Metab Care 14(1), 28–34. 10.1097/MCO.0b013e32834121b1.

[7] Giorgi, C., Marchi, S., Simones, I. C. M., Ren, Z., Morciano, G., Perrone, M., Patalas-Krawczyk, P., Borchard, S., Jedrak, P., Pierzynowska, K., Szymański, J., Wang, D. Q., Portincasa, P., Wegrzyn, G., Zischka, H., Dobrzyn, P., Bonora, M., Duszynski, J., Szabadkai, G., Zavan, B, Olivieira, P. J., Sardao, V. A., Pinton, P., and Wieckowski, M. R. (2018). Mitochondria and reactive oxygen species in aging and age-related diseases. Int Rev Cell Mol Biol 340, 209–344. 10.1016/bs.ircmb.2018.05.006.

[8] Zhang, K., Ma, Y., Luo, Y., Song, Y., Xiong, G., Ma, Y., Sun, X., and Kan, C. (2023). Metabolic diseases and healthy aging: identifying environmental and behavioral risk factors and promoting public health. Front Public Health 11, 1253506. 10.3389/fpubh.2023.1253506.

[9] Vijg, J. and Dong, X. (2020). Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. Cell 182(1), 12–23. 10.1016/j.cell.2020.06.024.

[10] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. Cell 153(6), 1194–1217. 10.1016/j.cell.2013.05.039.

[11] López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2023). Hallmarks of aging: An expanding universe. Cell 186(2), 243–278. 10.1016/j.cell.2022.11.001.

[12] Shokhirev, M. N., and Johnson, A. A. (2021). Modeling the human aging transcriptome across tissues, health status, and sex. Aging Cell 20(1), e13280. 10.1111/acel.13280.

[13] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst 1(6), 417–425. 10.1016/j.cels.2015.12.004.

[14] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., … and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences, 102(43), 15545-15550. https://doi.org/10.1073/pnas.0506580102

[15] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell systems, 1(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004