communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-08845-8

Human Cell Aging Transcriptome Atlas (HCATA): a single-cell atlas of age-associated transcriptomic alterations across human tissues

Josh Bartz [1,2], Xiao Ma [1,2], Lei Zhang[1,3] & Xiao Dong [1,2]

Biological aging is associated with progressively more severe genetic and epigenetic alterations. While these changes are expected to affect the transcriptional profile of cells, the magnitude of that effect is unknown as the aging transcriptome is still poorly understood. Understanding the aging transcriptional landscape will give us greater insight into how cells are affected by and/or respond to the aging process. To facilitate the large-scale exploration of the aging transcriptome, we report the development of the Human Cell Aging Transcriptome Atlas (HCATA). HCATA, contains single-cell RNA-sequencing datasets from 76 publications totaling 92 million cells and 3,475 tissue-level samples across more than 50 tissue types with ages ranging from 0 to 103 years. HCATA includes a genome browser that allows users to interactively explore age-related differential expression, as well as functions to explore related pathways at the tissue and cell-type level. HCATA is publicly accessible at http://hcata-xiaodonglab.org:3304.

Biological aging is characterized by widespread functional decline and an increased risk of injury, disease, and mortality[1–3]. While the pronounced damage to the genome and epigenome plays a crucial role in the aging process[4,5], a clear mechanism by which they might drive aging is unknown[3]. One purposed mechanism is through alterations in the transcriptional landscape[6,7]. Despite recent advances in RNA sequencing technology, especially at the single-cell level, the aging transcriptome is poorly understood and the ways in which it is affected by and/or responds to the aging process is unclear. A comprehensive profiling of the aging transcriptome at the single-cell level has been performed in mice[8,9] and drosophila[10], but doing so in humans has so far proven to be prohibitively expensive, because of the huge genetic and environmental variations in humans.

One way to address this issue is to develop a comprehensive data resource of existing single-cell RNA sequencing (scRNA-seq) data of different human tissues across a wide age range. Recently, a few repositories of human scRNA-seq data curated for aging research have been developed, such as AgeAnno[11], Aging Atlas[12], and SCAD-Brain[13]. However, these databases have a relatively small number of samples and alone are not sufficient to provide a complete picture of the aging transcriptional landscape. At the opposite extreme, large repositories of scRNA-seq data like GEO, SRA, and the Human Cell Atlas do not focus on a particular field of research, e.g., aging, and only provide raw data. Listing age-related studies from these databases is very time consuming, and further exploration and visualization of age-related changes from multiple studies directly from a database is almost impossible. Therefore, there is still an unmet need for a large database that allows users to profile the aging transcriptome in a wide range of cellular contexts.

To address this gap, we developed the Human Cell Aging Transcriptome Atlas (HCATA), which provides a comprehensive repository of human scRNA-seq data from various tissues spanning nearly the entire lifespan. Additionally, HCATA's online portal contains tools that allow users to explore age-related transcriptional changes including differential gene expression and Gene Ontology (GO) term enrichment.

Results and discussion

Database construction and content

To develop HCATA, we first manually curated the PubMed database for original studies that generated scRNA-seq data from human tissues. Studies were included if they meet the following requirements: containing publicly available age and tissue information for every sequenced

[1]Masonic Institute on the Biology of Aging and Metabolism, University of Minnesota, Minneapolis, MN, USA. [2]Department of Genetics, Cell Biology and Development, University of Minnesota, Minneapolis, MN, USA. [3]Department of Biochemistry, Molecular Biology, and Biophysics, University of Minnesota, Minneapolis, MN, USA. e-mail: dong0265@umn.edu

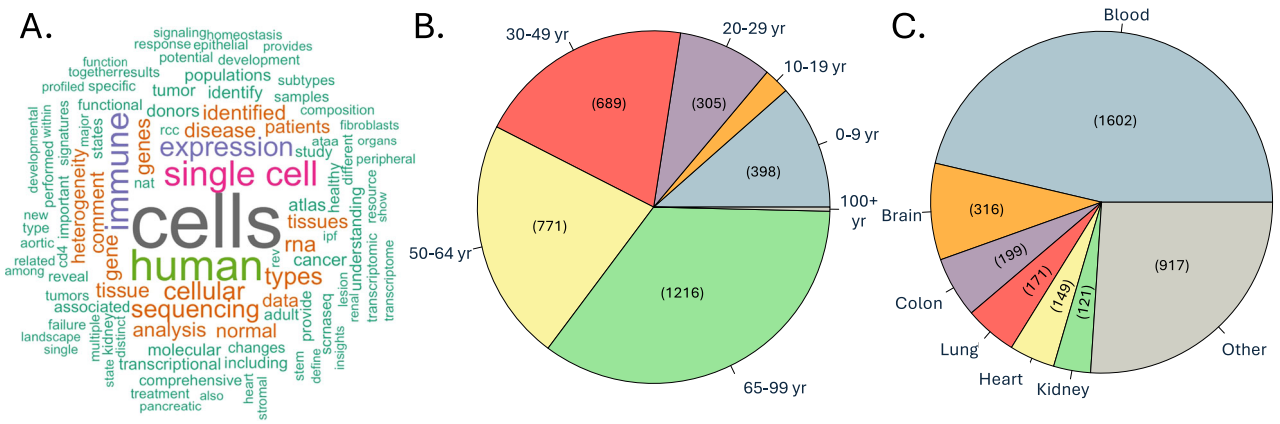

Fig. 1 | HCATA's data distribution. A Word cloud of the abstracts from the datasets included in HCATA. Pie charts showing the **B** age and **C** tissue distributions of the datasets.

sample; having a high proportion of samples taken from normal, non-diseased tissue; and being able to match its metadata with each sample's gene expression profile. In total, scRNA-seq data of 76 different studies were collected (Table S1), and Fig. 1A summarizes the most frequent words used in the abstracts of these studies. Both normal and diseased samples from the studies are included in our database to allow researchers to profile the aging transcriptome in multiple contexts. Together, our scRNA-seq database consists of 92 million cells and 3475 tissue-level samples across more than 50 tissue types, with ages ranging from 0 to 103 years (Fig. 1B, C). All datasets have been stored in a standard format to facilitate easy access, usage, and download.

To provide a user-friendly interface to this comprehensive data resource, we developed HCATA's database architecture by integrating two software components: a database engine and a database management system (DBMS) using InnoDB and MySQL, respectively. In brief, the InnoDB engine provides the core framework for storing, editing, and retrieving data, and is configured to balance data storage and query speed. MySQL manages queries and user interactions within the database, and enforces stored data to follow rigid structures with set relationships (Fig. 2).

HCATA's database component can be accessed online using an accompanying website created with Angular, a TypeScript based framework that leverages components and two-way data binding to build dynamic web applications. It communicates with our backend database through a secure Node.js Express server. This setup allows users to access and download thousands of standardized samples, and additionally, explore age-related differentially expressed genes (DEGs) using an interactive genome browser[14] as well as their enriched pathways (see below; Fig. 3A).

Differential gene expression and GO analyses

The major challenge in integrating scRNA-seq datasets from different literature sources is their heterogeneity in the usage of experimental platforms and data processing pipelines. To minimize the noise caused by varying experimental platforms, we focused the analyses of DEGs from data derived from the Chromium 10x platform, which is the most widely used platform in the studies in our collection. A few additional criteria were applied: (i) inclusion of at least five normal samples, (ii) presence of at least one sample from an individual over 60 years of age, and (iii) a minimum age span of 25 years between the youngest and oldest samples. Although the above limits our statistical power, it avoids introducing severe batch effects. Overall, 287 tissue samples from 11 selected studies were selected for the following analyses (Table S2).

To address the challenge in data processing, we developed a standard data processing pipeline, which was applied separately to each selected dataset. This pipeline includes two major components: data preprocessing and the analysis of age associations (Fig. 3B). For data preprocessing, the pipeline is based on the R package "Seurat"[15], including five key steps: cleaning and quality control, normalization, integration, clustering, and cell type annotation (Methods). Clusters were annotated based on their gene expression profiles. For each cluster, the three most probable annotations were identified, with the highest-confidence label designated as the primary annotation, followed by the secondary and tertiary annotations. The results of clustering and cell type annotation were validated by manually comparing the corresponding results from each of the 11 studies selected above. To further validated our approach, we compared the original clustering and annotation of a subset of data from the Tabula Sapiens[16], to the results generated by our standard pipeline. In this dataset, we found a high concordance: our standard clustering and annotations agreed with the original study for approximately 72% of the 34,593 cells that passed both ours and their filters (Fig. S1 and Table S3). For 24% of cells, while the primary annotation did not match, the secondary annotation agreed with the original study. In total HCATA correctly annotated 96% of cells when considering both primary and secondary annotations. The remaining 4% discrepancy was attributable to minor differences in clustering and markers used for annotation. The above demonstrates that our approach reliably approximates clustering and annotation of the original study.

To test age associations, we identified age-related DEGs at both the cell- and tissue-type levels for each cell cluster and each dataset separately. We did not integrate raw data together because data integration is highly affected by batch effects, and instead we summarized age-related association for each gene across different cell clusters and studies through data visualization in the web interface and the usage of the Gini Coefficient for pathways (described below). Unlike most previous studies, HCATA calculates DEGs by using age as a continuous variable instead of grouping samples into young, middle, and/or old categories. This provides a more unbiased measurement of age-related changes in gene expression. One other challenge here is that conventional methods, e.g., "Seurat", are not designed to analyze differences in scRNA-seq across multiple subjects. We used the R package "NEBULA: Negative Binomial mixed model Using a Large-sample Approximation"[17], which addresses both challenges. Using Nebula, we calculated DEGs for each cell cluster within samples in each study using age as a continuous variable and sex as a covariate. To validate NEBULA's performance, we selected the significant DEGs and GO Terms (see below) from three studies and compared them with findings in the literature (Tables S4 and S5).

Next, we performed a Gene Set Enrichment Analysis (GSEA) to discover the age-associated Gene Ontology (GO) terms using clusterProfiler[18]. Gini Coefficient was then estimated for each GO term to quantify their similarity in age-related changes across different cell types, tissue types, and studies. Based on the DEGs above, we present the results from the 11 studies across 8 tissues and 55 cell types. In Figs. 4A and 5A, the top 5 most significantly enriched GO terms in each tissue and cell type, respectively, along with their enrichment levels in the other tissues and cell types are

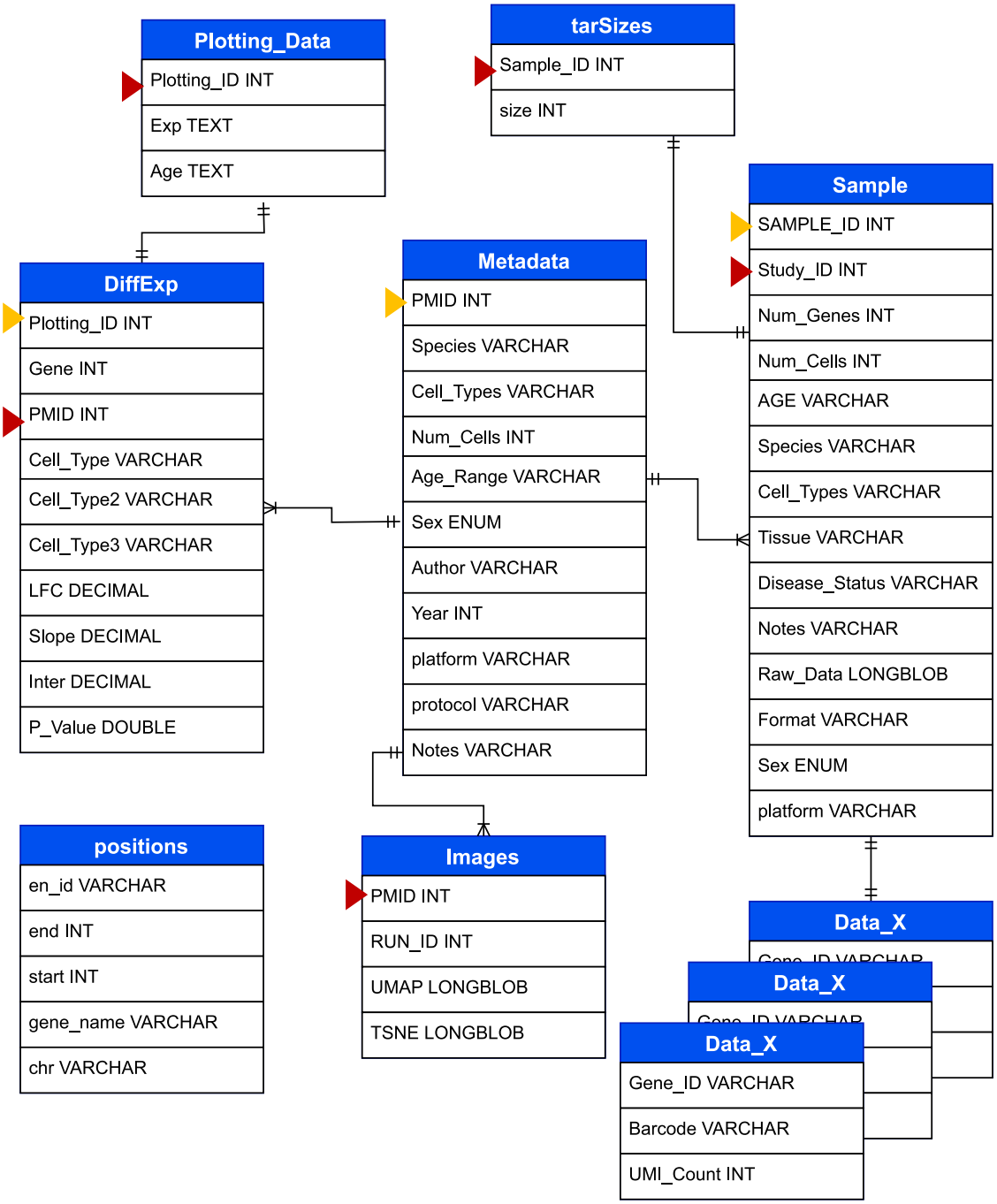

Fig. 2 | Entity-relationship diagram of HCATA. The Entity-Relationship Diagram for HCATA represents the conceptual framework of the database structure, outlining the key entities, their attributes, and the relationships between them. Primary keys used to uniquely identify entries are marked in yellow, while foreign keys that help to connect tables are marked in red. Related tables are connected by lines that indicate the nature of their relationships. A line connecting two tables shows whether the relationship is one-to-one or one-to-many. In a one-to-one relationship, each entry in one table corresponds to exactly one entry in the related table. Conversely, in a one-to-many relationship, a single entry in one table can be associated with multiple entries in the related table.

presented. Based on the Gini Coefficients, both universally and specifically enriched GO terms were identified, with many notably related to inflammation (Figs. 4B, C and 5B, C). While age-related chronic inflammation, or "inflammaging", is a hallmark of aging that has been detected in many tissues and cellular contexts[3,19–21], its relationship with the aging transcriptome remains largely unexplored[22,23]. The universally enriched GO terms that we found underlying inflammaging include the response to bacterium, inflammatory response, and chemotaxis pathways. These pathways likely represent the classical signs of inflammaging and lead to increased and pervasive expression of multiple pro-inflammatory markers. We also detected inflammaging related GO terms that are tissue and cell type specific; for instance, immunoglobulin production is primarily upregulated in bone marrow, while changes in the immune response pathway are mostly observed in the brain, liver, and skeletal muscle (Fig. 4A). Pathways associated with cilia, which are hypothesized to regulate immune response in a variety of cell-types[24–27], are significantly downregulated in macrophages compared to other analyzed cell types (Fig. 5A). These results suggest that age-related alterations inhibit cilium function in macrophages, potentially exacerbating age-related declines in immune function. Additionally, multiple inflammaging pathways are enriched for both upregulated

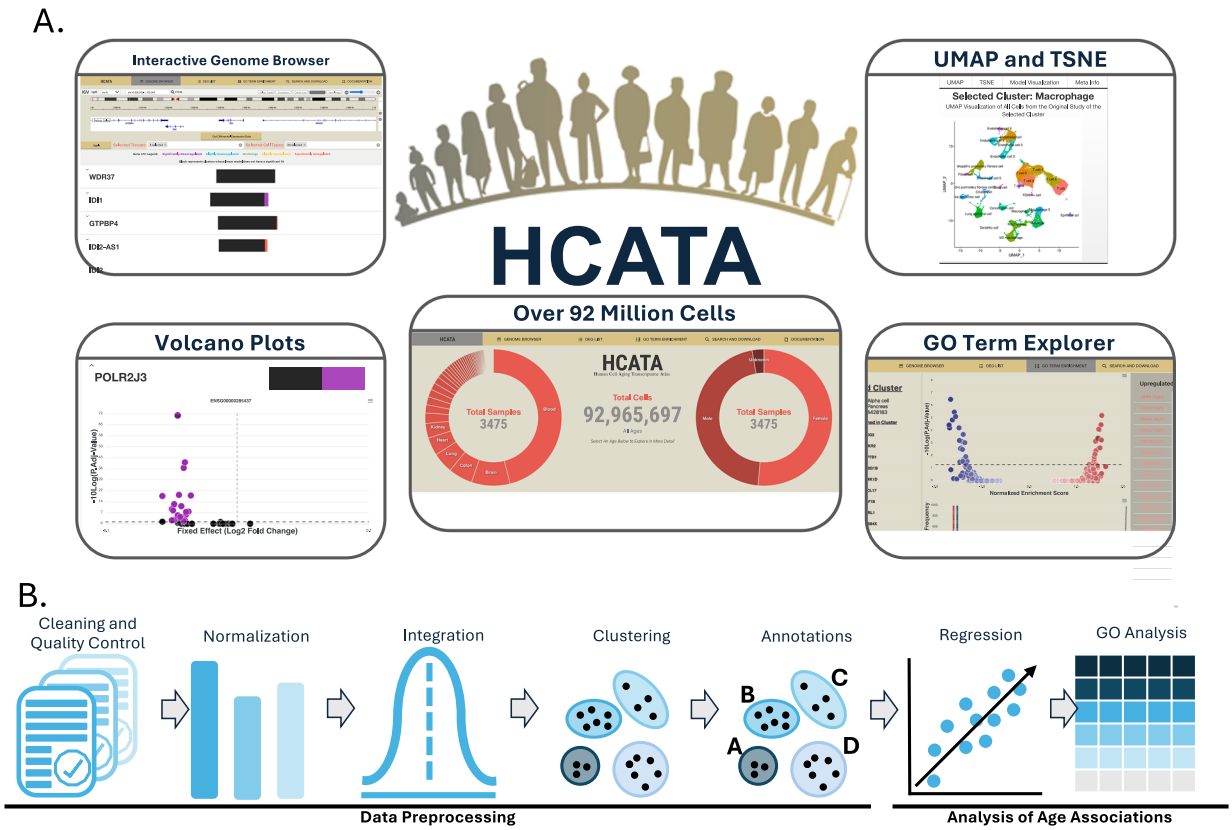

Fig. 3 | HCATA's web functions. A HCATA provides a user-friendly website for data visualization to explore results of cell clustering, DEGs, and GSEA. **B** Key steps in HCTAT's standard pipeline used to process and analyze the scRNA-seq data.

and downregulated genes, suggesting that the aging transcriptome encoding inflammation and immune function undergoes significant and complex changes with age. Unlike inflammaging, all pathways associated with cell differentiation and development were tissue or cell type specific (Figs. 4A and 5A) and only enriched in upregulated genes (Figs. 4B and 5B). Other enriched pathways that are known to have age-related alterations include the electron transport chain pathways, TORC1 signaling, cell adhesion, and angiogenesis. Additionally, enrichment at the cell-type level, does not always mirror enrichment at the tissue level, suggesting that age-related transcriptional changes can be highly cell-type specific and may be obscured in bulk tissue analysis. For example, cytoplasmic translation is strongly upregulated in the pancreas at the tissue level (Fig. 4A), but moderately downregulated in alpha cells (Fig. 5A).

To further investigate age-related pathways that had been previously defined in literature we compiled a collection age-related GO terms enriched by genes reported in the GenAge database[28]. We found that at the tissue level, numerous age-related pathways exhibit either positive or negative enrichment (Fig. 6A). However, at the cell-type level fewer pathways, especially downregulated pathways, show a statistically significant change, likely reflecting the reduced statistical power associated with analyzing a single cell type (Fig. 6B). Overall, in addition to providing scRNA-seq data and DEGs with age, HCATA allows its users to explore age-related transcriptional changes that are both general and tissue or cell-type specific to the aging process.

Web interface overview

HCATA provides a web interface through which users can retrieve scRNA-seq datasets from over 3000 samples and explore age related transcriptional changes. Our web interface is separated into 6 pages: Home, Genome Browser, DEG List, GO Term Enrichment, Search and Download, and Documentation. The Home page serves as the landing page and displays a graphical overview of the database.

The Genome Browser page provides users with an interactive platform to explore the DEGs that arise during normal human aging (Fig. 7A). An imbedded Integrative Genomics Viewer (IGV) browser allows users to explore different areas of the human genome using simple point and click mechanics. Once a region of interest has been selected using IGV, age related transcriptional changes for the genes within the selected region are displayed. These changes are summarized for each gene in a volcano plot, which displays differential expression for a given gene in all cell clusters across the 11 studies that we used. In the volcano plots, all data points have been colored based on statistical and biological significance. Because differential expression is calculated using age as a continuous variable, we use the Log2FC (log2 fold-change) per year in the volcano plots to present age-related change in gene expression. This can be unintuitive to interpret as a yearly Log2FC of 0.0116 represents the doubling of a gene's expression over 60 years. To make these numbers easier to interpret, we include an "Average Change in Expression per 10 Years" metric that converts this annual Log2FC to a percent change every decade. This metric can be found by clicking a data point (i.e., a cell cluster in a study) in a volcano plot and under the "Meta Info" tab in the figures shown on the right of the screen. Other information is also shown including associated UMAP (Uniform Manifold Approximation and Projection) reduction, t-SNE (t-distributed Stochastic Neighbor Embedding) reduction, alternate cell type annotations, and other information about the clusters in the original study.

DEGs can also be browsed based on their prevalence across tissues and cell types in the DEG List page. This page lists all DEGs ordered by the number of cell clusters in which they showed a statistically significant change with age, i.e., defined by an adjusted $P < 0.05$ and a Log2FC change ≥0.0116 (see the above).

The GO Enrichment page serves as a gateway for users to explore both positively and negatively enriched pathways related to normal human aging (Fig. 7B). Users can select a pathway of interest, and HCATA will display the

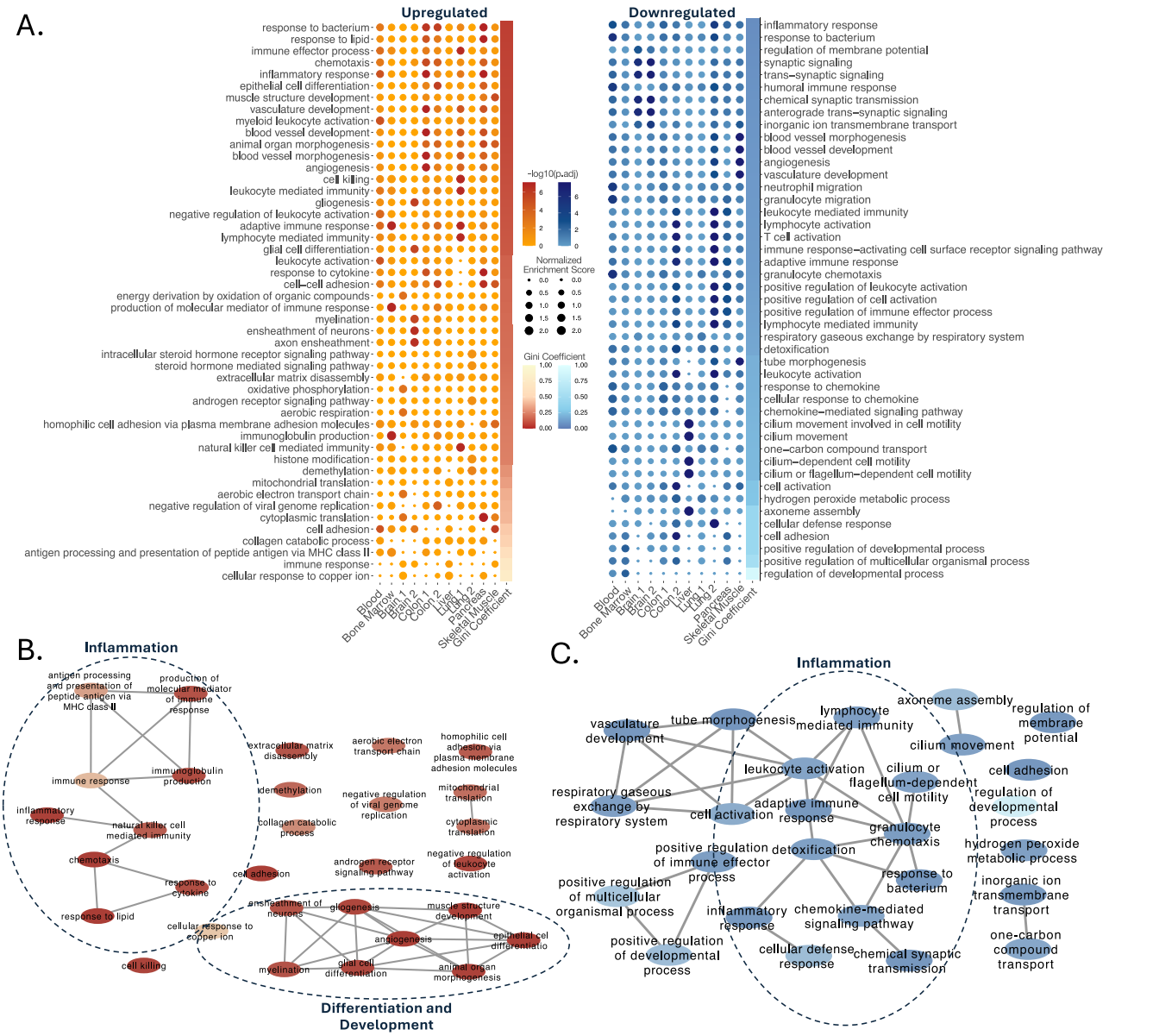

Fig. 4 | Gene set enrichment analysis of GO terms related to aging at the tissue level. A Dotplots of the top 5 most significantly enriched term for each tissue are shown for upregulated (red) and downregulated (blue) genes. Terms are ordered by Gini Coefficient with most universally enriched terms on the top. **B**, **C** GO Term visualizations were performed using REVIGO[36]. GO Terms are color-coded based on their Gini Coefficient and are connected to other terms based on the pairwise similarities. Several inflammaging pathways are significantly enriched among both **B** upregulated and **C** downregulated genes and contain universal and specific terms.

enrichment scores for that pathway across multiple cellular contexts. These scores are summarized in a color-coded volcano plot, which presents age-related changes of a selected pathway in all cell clusters across the 11 studies (Fig. 7B, *center panel*). By default, HCATA shows enrichment scores for all cell types and tissue types, but users can opt to display a specific subset of results. Within the selected cell types, genes associated with the selected pathway are ranked based on their consistent upregulation or down-regulation (Fig. 7B, *right column*), helping users identify genes that consistently drive enrichment across different cellular contexts. Like the Genome Browser page, users can select clusters of interest to view more detailed information about those clusters and the specific genes enriched within them (Fig. 7B, *left column*). Finally, the Gini Coefficient for the selected pathway is shown to demonstrate its cell-type specificity of enrichment.

The Search and Download page allows users to download raw scRNA-seq data matrices and their associated metadata from HCATA (Fig. 7C). Users can specify their desired dataset(s) using a query system based on their search criteria. HCATA contains over 3000 samples, so many query results may return hundreds if not thousands of samples. This amount of data can be hard to visualize as a list, so the Search and Download page also includes a series of pie charts that summarize the health, sex, age, and tissue distribution of samples that match the users search criteria. Users can select multiple samples and download them all at once, making it easier to curate large datasets.

Overall, the web component of HCATA was designed for users to visualize DEGs and enriched pathways in customizable figures. These visualization tools make it much easier to profile the aging transcriptome compared currently existing web tools that only display results as a series of tables.

Limitations

A few limitations exist. First, it is difficult to integrate multiple scRNA-seq datasets together, largely due to technical variability and batch effects inherent to scRNA-seq data. To address this, we performed analyses

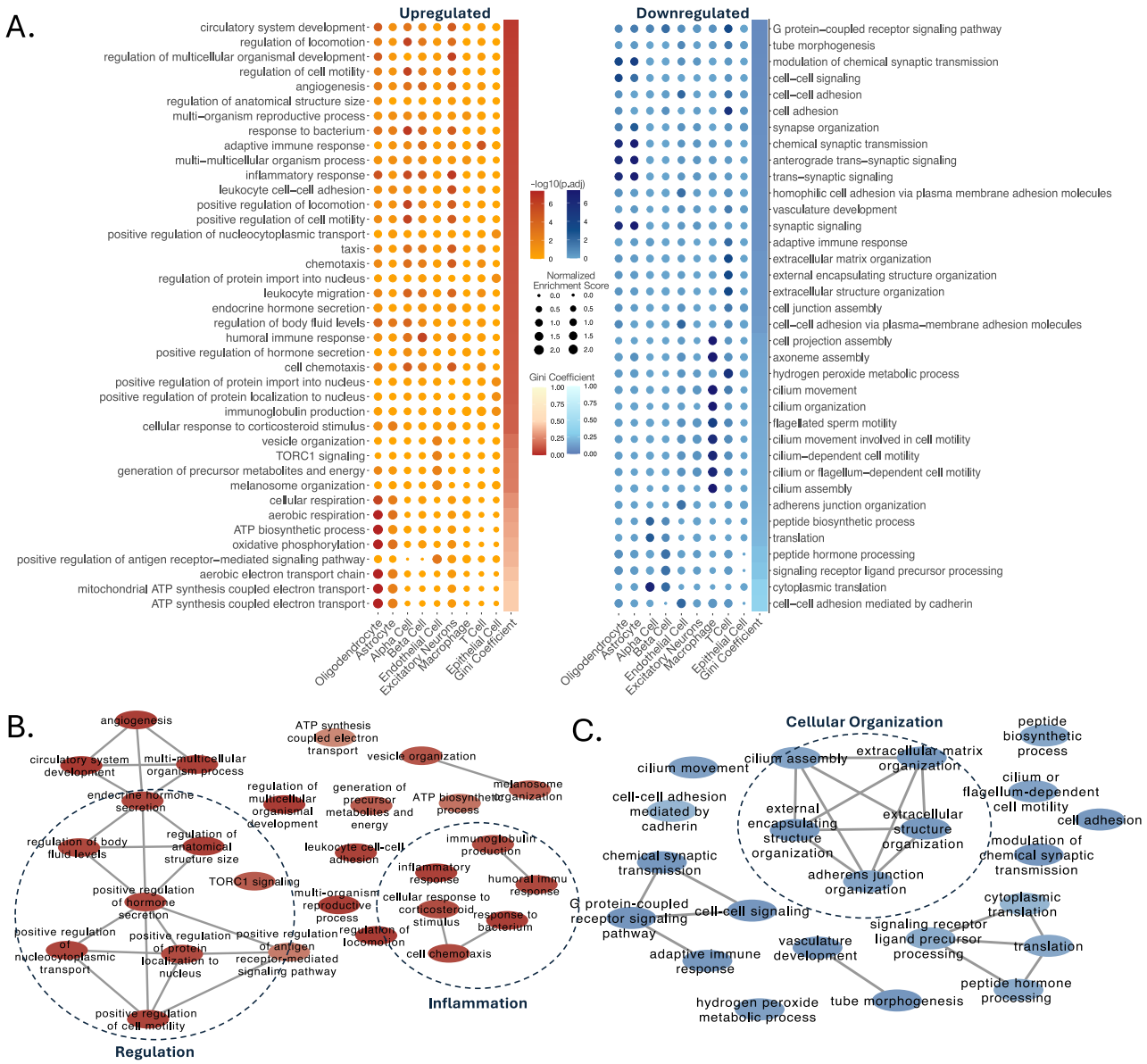

Fig. 5 | Gene set enrichment analysis of GO terms related to aging at the cell-type level. A Dotplots of significantly enriched terms are shown for upregulated (red) and downregulated (blue) genes. The top 5 most significantly enriched GO terms for each cell type are displayed. Terms are ordered by Gini Coefficient with most universally enriched terms on the top. GO Term visualizations were performed using REVIGO[36] for **B** upregulated and **C** downregulated terms. GO Terms are color-coded based on their Gini Coefficient and connected to other terms based on the pairwise similarities.

separately on each dataset and later summarized results, which limits the statistical power of DEG and GO analysis. Second, developing a universal pipeline for data analysis is challenging, because some parameters, e.g., doublet rate, are dataset specific and can vary significantly across experimental platforms. To address this, we applied a balanced cutoff suitable for a wide variety of studies and restricted our analysis to datasets generated on the 10x Chromium platform. Finally, an absence of environmental metadata limited our ability to account for environmental factors that may influence the transcriptome when identifying DEGs.

Conclusion

In conclusion, HCATA offers the aging research community a comprehensive and standardized repository of scRNA-seq data spanning a wide range of tissues and age groups. HCATA currently contains 3475 human samples, and future updates will continue to expand the database and its function. Compared to similar existing databases[11–13], HCATA contains significantly more datasets and visualization tools, enhancing its utility in exploring the aging transcriptome. Leveraging this extensive dataset, we identified DEGs associated with normal human aging across multiple tissues and cell types. Our analysis suggests that inflammation-related pathways are consistently enriched with age, with both specific and non-specific pathways emerging across different tissues and cellular contexts. These results demonstrate HCATA as a powerful resource for advancing our understanding of the aging transcriptome.

Methods

Database construction

HCATA was constructed by integrating two software components: an engine and a DBMS. A database engine is the underlying software component that determines how data is stored to balance efficient memory usage with efficient query speed. A DBMS system acts as a conduit for human users to easily interact with the underlying data structure that the engine has creased. It has three primary uses, enforcing database relationships, executing queries, and managing users. HCATA uses InnoDB (version

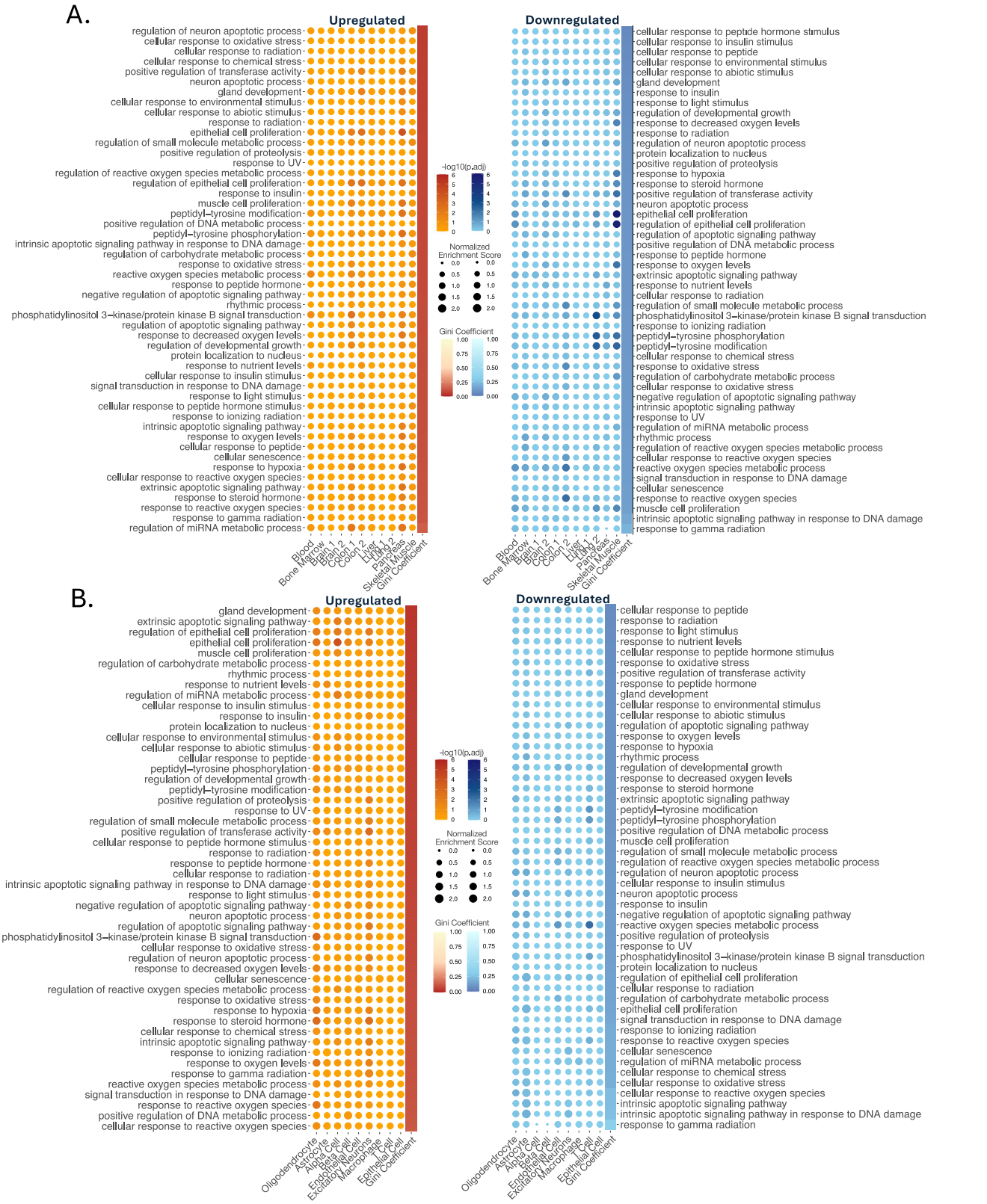

Fig. 6 | Gene set enrichment analysis of aging GO terms. A Dotplots of the 50 most significantly enriched GO terms in the genes from the GenAge database at the tissue level, shown for upregulated (red) and downregulated (blue). Terms are ranked by Gini Coefficient with the most universally enriched terms on the top. **B** Dotplots of the 50 most significant GO terms enriched in the genes from the GenAge database at the cell-type level, shown for upregulated (red) and downregulated (blue). Terms are ranked by Gini Coefficent with the most universally enriched terms on the top.

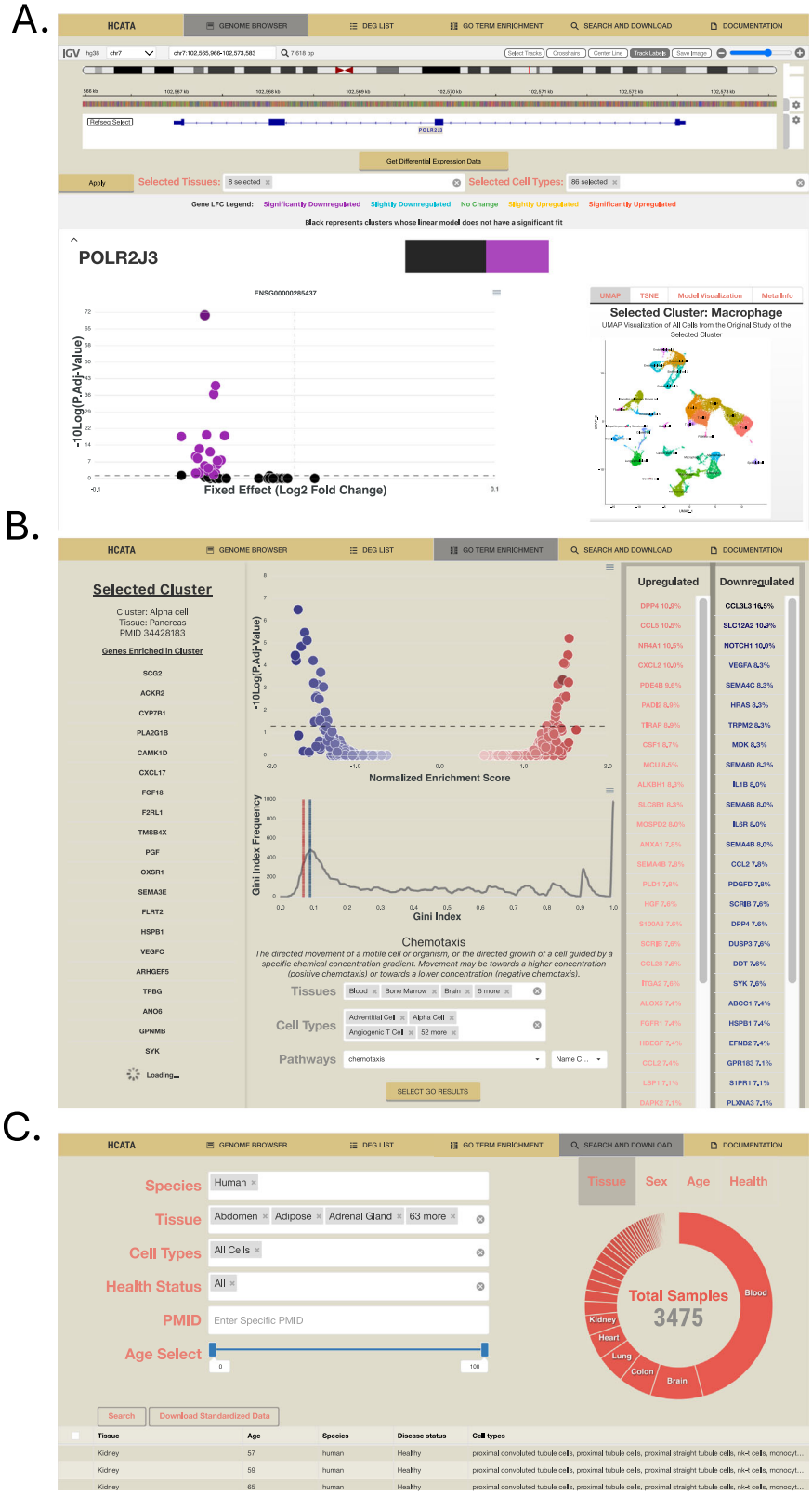

Fig. 7 | Tools included in HCATA's web interface. **A** The Genome Browser page allows users to explore age-related changes to the transcriptome across cell types and tissues. **B** The GO Enrichment page allows users to explore pathways that are positively or negatively enriched with age along with the genes that drive enrichment. **C** The Search and Downloads page allows downloading scRNA-seq datasets from HCATA.

8.0.3) as its engine and MySQL (version 8.0.3) as its DBMS. To handle the large datasets in HCATA, its database engine was configured balance data storage and query speed. MySQL was utilized to enforce a semi-relational database using the structure outlined in Fig. 2. This set structure makes it easy to create complex queries that aggerate data from different tables into a single output. To ensure security, MySQL also manages user specific read/write access into and out of the database.

In addition, HCATA includes an online portal that allows users to interact and download data from out database. This website was built using the Angular Framework (version 13.1.4) based on TypeScript (version 4.5.1). Angular allows users to build dynamic single page website by utilizing two-way databinding which dynamically updates the webapp based on client-side inputs. Our website utilizes the IGV (version 2.13.5)[14,29] to allow users to interactively explore age-related DEGs and GO Terms. Angular will

retrieve data from our backend database by triggering HTTP requests that are handled by a Node.js Express server (version 4.19.2). Node.js Express is a lightweight version of Node.js that handles HTTP requests triggered by Angular. Specifically, the Node.js Express server will translate HTTP requests into SQL queries, execute those queries on the database, and return the resulting data to the Angular frontend.

Dataset standardization

Raw scRNA-seq datasets and metadata files were downloaded from their original databases in various formats. To better facilitate the large-scale exploration of the aging transcriptome, datasets in HCATA have undergone a standardization process. Datasets were annotated with gene labels using Ensembl IDs obtained through Ensembl BioMart[30]. Gene Names or HGNC IDs that cannot uniquely match to an Ensembl ID were removed. Raw scRNA-seq count matrices were stored in a matrix mart file format which consists of three files: a barcode file, feature file, and matrix file. A barcode file contains the unique names of each cell while a feature file contains unique gene names. A matrix file is a sparse matrix file that leverages scRNA-seqs high dropout rate to efficiently store data. Since traditional matrices would be inefficient due to the predominance of zero values, sparse matrices address this issue by representing only the non-zero values. Despite matrix mart format consisting of three files, it is the most space efficient way to store scRNA-seq data. Metadata for all samples are stored in a lightweight easy to read table, and both metadata and count matrices are split up by sample to allow users greater control when downloading data.

Data preprocessing

In preparation for further analysis, selected datasets underwent five preprocessing steps including cleaning and quality control, normalization, integration, clustering, and cell annotation (Fig. 3B). To minimize bias that can arise during the preprocessing steps, each study underwent the same preprocessing steps using the same parameters. *Cleaning and quality control:* We removed potential doublets and low-quality cells whose total UMI counts were in the top or bottom five percent of the all the cells within a dataset. These cutoffs strike a balance between permissiveness and rigor, recognizing that multiplet rates vary across datasets and are challenging to generalize. While it is impossible to eliminate all multiplets without mistakenly discarding single cells, a 5% cutoff effectively removes most multiplets while preserving the majority of valid cells. Each dataset underwent normalization and clustering based on the standard Seurat pipeline[15,31,32]. *Normalization:* UMI counts in each cell are divided by the library size of that cell, then multiplied by a scale factor of 10,000, and finally transformed using the natural logarithm to produce normalized expression values. *Integration*: Seurat first identifies the most variable genes across all cells within a dataset, as these genes capture key biological differences between cells. Using this shared gene set, pairs of cells with similar expression profiles are then selected as "anchors" to guide integration. Finally, samples are integrated together using their shared anchors to correct for technical variation. *Clustering*: For each study, UMAP and t-SNE reduction were derived and included within HCATA. These reductions also proved useful in validating our preprocessing pipeline as our reductions should match up with the reductions generated by the original studies. *Cell annotation*: Clusters were assigned a cell type based on the degree of overlap between genes uniquely expressed with a cluster and marker genes from Cell Marker[33]. A limitation is imbalances between the number of total markers for given cell types. In brief, cell types with more associated markers have a higher chance of overlapping with a cluster than other cell types with fewer markers. To account for this bias, we provided each cluster two additional possible cell types in addition to its most likely cell type.

We manually checked each of our reductions against the original paper to ensure that a similar distribution of clusters and cell types were detected. As an additional validation step, we directly compared the clustering and annotation of cells from a subset of the Tabula Sapiens[16] against HCATA's clustering and annotation (Fig. S1 and Table S3). To assess the level of agreement between this study and HCATA, we assume that Tabula Sapiens' "Hematopoietic Multipotent Progenitors" correspond to HCATA's "Hematopoietic Stem Cells," and that Tabula Sapiens' "Hematopoietic Precursor Cells" are equivalent to HCATA's "Myeloid" cells. Additionally, HCATA identifies a population of "Lymphoid-Primed Multipotent Progenitors," which represent a subset of hematopoietic multipotent progenitors that have not yet differentiated into precursor cells. As such, these may correspond to either Tabula Sapiens' "Hematopoietic Multipotent Progenitors" or "Hematopoietic Precursor Cells." Unfortunately, we could not validate other studies in this way as few studies provided the clustering annotations for individual cells in their publication.

This standard processing pipeline is tailored to fit data generated by the 10x Genomics platform. Other platforms, e.g., Smart-Seq2, have different technical limitations, which could lead to bias when applying the same pipeline to multiple platforms. To control for this bias, we limited our analyses only on datasets generated with the 10x Genomics platform.

DEG and GO term enrichment analyses

To detect DEGs associating with age, we used NEBULA[17]. NEBULA (NEgative Binomial mixed model Using a Large-sample Approximation) is an algorithm used to detect DEGs in scRNA-seq data. Nebula can calculate changes across continuous dependent variables like age, while many other models that can only handle categorical ones. Additionally, NEBULA is optimized to detect DEGs in studies across many samples. While a larger sample size is often important for statistical analysis, it can bias models to detect spurious DEGs that arise randomly due to overdispersions between samples. NEBULA uses negative binomial mixed models to independently assess the overdispersion arising from samples and cells, which greatly reduces false positive DEGs. While negative binomial mixed models are ideal for use in scRNA-seq data, they are computationally intensive. NEBULA's algorithms greatly increase the computation speed of negative binomial mixed models, allowing for their use in a reasonable amount of time.

NEBULA was used to calculate DEGs for each cluster of cells within every study using age as a continuous variable and sex as a covariate. To ensure the detected DEGs to be age related, we calculated DEGs for each cluster of cells individually to avoid detecting DEGs that arise between different cell types. Additionally, DEGs were only calculated in normal human samples to avoid disease-related DEGs. However, we were unable to control for DEGs that may arise due to environmental differences between samples, because most published scRNA-seq data lacks the appropriate metadata needed to control for these effects. To validate NEBULA's detection of age-related genes we compared the twenty most significantly changing genes in three studies to the literature (Table S4).

DEGs calculated by NEBULA were then used to detect enriched Biological Process (BP) GO Terms. GO enrichment was done separately for up and downregulated genes using the clusterProfiler package[18]. To validate GO Term analysis the top 10 positively and negatively enriched GO Terms from three studies were compared to the literature (Table S5). To select GO Terms to display for visualization (Figs. 4 and 5) we selected the top five most enriched pathways from each group. For example, in Fig. 4 the top 5 upregulated enriched pathways for the Brain_2 enriched group, including *gliogenisis, glial cell differentiation, myelination, ensheathment of neurons, and axon ensheathment*, are included in the figure.

For Fig. 6, to define aging-related pathways, we used all 302 human genes cataloged in the GenAge database[28] and identified associated GO terms using Over Representation Analysis (ORA). ORA identified 325 GO terms that were strongly association (adjusted $P < 1e\text{-}10$) with the GenAge genes. We selected the 50 pathways with the most significant enrichment as indicated by the lowest adjusted P values.

Tissue-specific and non-specific GO terms with age

To quantify tissue and cell-type specificity, we used the Gini Coefficient, which is a classical mathematical measurement of uniqueness (Eq. 1). There

are many forms

$$Gini = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n^2\bar{x}}$$
$$n = number\ of\ observations,\ x_{i/j} = value\ of\ individual\ observation \quad (1)$$

of the Gini Coefficient; we use the Pairwise Gini Coefficient, which is defined as normalized average difference between all observations[34,35]. A Gini Coefficient is a number between 0 and 1, where 1 represents total uniqueness and 0 represents no uniqueness. GO terms with low Gini indices are enriched nearly equally across groups, while those with high Gini indices are enriched in only a few. To calculate the uniqueness of a GO Term within our groups we treat the Normalized Enrichment Scores for each group as an observation. Therefore, x_i represented the enrichment score for a group (i.e., Blood, Oligodendrocyte).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Statistics and reproducibility

Statistical tests for identifying DEGs were included in Nebula[17], while tests for identifying significantly enriched GO Terms was included in clusterProfiler[18]. The number of samples analyzed in each study are shown in Supplementary Table S2.

Data availability

All data is available for download and exploration at http://hcata-xiaodonglab.org:3304. Further information can be found in the Supplementary Information file and source data for all created figures can be found in Supplementary Data 1.

Received: 6 October 2024; Accepted: 4 September 2025;
Published online: 09 October 2025

References

1. Niccoli, T. & Partridge, L. Ageing as a risk factor for disease. *Curr. Biol.* **22**, R741–R752 (2012).
2. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
3. Lopez-Otin, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. Hallmarks of aging: an expanding universe. *Cell* **186**, 243–278 (2023).
4. Soto-Palma, C., Niedernhofer, L. J., Faulk, C. D. & Dong, X. Epigenetics, DNA damage, and aging. *J. Clin. Investig.* **132**, https://doi.org/10.1172/JCI158446 (2022).
5. Yousefzadeh, M. et al. DNA damage-how and why we age? *Elife* **10**, https://doi.org/10.7554/eLife.62852 (2021).
6. Bartz, J., Jung, H., Wasiluk, K., Zhang, L. & Dong, X. Progress in discovering transcriptional noise in aging. *Int. J. Mol. Sci.* **24**, https://doi.org/10.3390/ijms24043701 (2023).
7. Perez-Gomez, A., Buxbaum, J. N. & Petrascheck, M. The aging transcriptome: read between the lines. *Curr. Opin. Neurobiol.* **63**, 170–175 (2020).
8. Tabula Muris, C. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature* **562**, 367–372 (2018).
9. Tabula Muris, C. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
10. Lu, T. C. et al. Aging Fly Cell Atlas identifies exhaustive aging features at cellular resolution. *Science* **380**, eadg0934 (2023).
11. Huang, K. et al. AgeAnno: a knowledgebase of single-cell annotation of aging in human. *Nucleic Acids Res.* **51**, D805–D815 (2023).
12. Aging Atlas, C. Aging Atlas: a multi-omics database for aging biology. *Nucleic Acids Res.* **49**, D825–D830 (2021).
13. Li, X. W. et al. SCAD-Brain: a public database of single cell RNA-seq data in human and mouse brains with Alzheimer's disease. *Front. Aging Neurosci.* **15**, 1157792 (2023).
14. Robinson, J. T., Thorvaldsdottir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics* **39**, https://doi.org/10.1093/bioinformatics/btac830 (2023).
15. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
16. Tabula Sapiens, C. et al. The Tabula Sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
17. He, L. et al. NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).
18. Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
19. Ferrucci, L. & Fabbri, E. Inflammageing: chronic inflammation in ageing, cardiovascular disease, and frailty. *Nat. Rev. Cardiol.* **15**, 505–522 (2018).
20. Franceschi, C. et al. Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann. N. Y. Acad. Sci.* **908**, 244–254 (2000).
21. Gerli, R. et al. Chemokines, sTNF-Rs and sCD30 serum levels in healthy aged people and centenarians. *Mech. Ageing Dev.* **121**, 37–46 (2000).
22. Rasa, S. M. M. et al. Inflammaging is driven by upregulation of innate immune receptors and systemic interferon signaling and is ameliorated by dietary restriction. *Cell Rep.* **39**, 111017 (2022).
23. Cheng, H. et al. Repression of human and mouse brain inflammaging transcriptome by broad gene-body histone hyperacetylation. *Proc. Natl. Acad. Sci. USA* **115**, 7611–7616 (2018).
24. Picon-Galindo, E., Latz, E. & Wachten, D. Primary cilia and their effects on immune cell functions and metabolism: a model. *Trends Immunol.* **43**, 366–378 (2022).
25. Kuek, L. E. & Lee, R. J. First contact: the role of respiratory cilia in host-pathogen interactions in the airways. *Am. J. Physiol. Lung Cell Mol. Physiol.* **319**, L603–L619 (2020).
26. Cassioli, C. & Baldari, C. T. A ciliary view of the immunological synapse. *Cells* **8**, https://doi.org/10.3390/cells8080789 (2019).
27. Baek, H. et al. Primary cilia modulate TLR4-mediated inflammatory responses in hippocampal neurons. *J. Neuroinflammation* **14**, 189 (2017).
28. de Magalhaes, J. P. & Toussaint, O. GenAge: a genomic and proteomic network map of human ageing. *FEBS Lett.* **571**, 243–247 (2004).
29. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
30. Dyer, S. C. et al. Ensembl 2025. *Nucleic Acids Res.* **53**, D948–d957 (2025).
31. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e3529 (2021).
32. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 e1821 (2019).
33. Zhang, X. et al. CellMarker: a manually curated resource of cell markers in human and mouse. *Nucleic Acids Res.* **47**, D721–D728 (2019).
34. Gini, C. Variabilità e Mutabilità (Tipogr. di P. Cuppini, Bologna, Italy, 1912).
35. Yitzhaki, S. & Schechtman, E. *The Gini Methodology: A Primer on a Statistical Methodology* (Springer, 2013).
36. Supek, F., Bosnjak, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE* **6**, e21800 (2011).

Acknowledgements
The authors would like to thank Tony Ni and Amir Haji for helping collect literature. This work was supported by NIH grants R00 AG056656 (X.D.), U19 AG056278 (X.D.), P01 AI172501 (X.D.), U54 AG076041 (X.D.), U54 AG079754 (X.D.), R35 GM159832 (L.Z.), and T32 AG029796 (J.B.); the Fesler-Lampert Chair for Aging Studies at the University of Minnesota (X.D.); and Sagol Network GerOmic Award for Junior Faculty from the American Federation for Aging Research (L.Z.).

Author contributions
X.D. and L.Z. conceived the study. J.B. developed the database, analyzed data, and wrote the first draft manuscript. X.D., L.Z., J.B., and X.M. contributed to editing the manuscript and provided critical feedback throughout the project.

Competing interests
L.Z. and X.D. are co-founders and shareholders of SingulOmics Corp. (New York, NY, USA). The other authors declare no conflict of interest.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-08845-8.

Correspondence and requests for materials should be addressed to Xiao Dong.

Peer review information *Communications Biology* thanks Chenxu Zhu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Rosie Bunton-Stasyshyn.