
**PONDICHERRY UNIVERSITY**
Puducherry – 605 014

# Identification of Anti-Ageing Biomarkers in Humans Using Transcriptomic Data

A Dissertation Submitted in Partial Fulfilment of the Requirements
for the Award of the Degree of

**Master of Science (M.Sc.) in Bioinformatics**

**Submitted by**
**Ankur Kumar**
Registration Number: 24MSBINPY0040

**Under the Supervision of**
Dr. Ayaluru Murali
Assistant Professor
Department of Bioinformatics

Department of Bioinformatics
Pondicherry University
Academic Year: 2024–2026

# Abstract

Ageing is a complex biological process characterised by progressive functional decline and increased susceptibility to disease. Identifying molecular biomarkers associated with ageing is essential for understanding its underlying mechanisms and developing potential therapeutic interventions. In this study, a comprehensive computational analysis was performed to identify age-associated transcriptomic signatures in humans using publicly available RNA-sequencing data.

Publicly available human transcriptomic data were obtained from the NCBI Gene Expression Omnibus (GEO) database and analysed using a reproducible Python-based bioinformatics pipeline. Gene expression data from primary human skin fibroblasts spanning a wide age range were processed, normalised, and subjected to differential expression analysis comparing young and aged samples. Multivariate analysis, gene-level visualisation, and functional enrichment analysis were employed to interpret age-associated transcriptional changes.

Principal Component Analysis revealed partial separation between young and aged samples, indicating global transcriptomic differences associated with ageing. Differential expression analysis identified a subset of genes involved in inflammatory signalling, developmental regulation, and tissue remodelling that exhibited consistent age-associated expression changes. Functional enrichment analysis further highlighted biological processes related to multicellular organism development, morphogenesis, and vascular development.

Overall, this study demonstrates that transcriptomic profiling provides a systems-level understanding of age-associated molecular alterations and enables the identification of potential biomarkers of ageing. The findings contribute to ongoing efforts to characterise the molecular basis of human ageing and support the use of computational approaches in ageing research.

**Keywords:** Ageing, Transcriptomics, RNA-seq, Bioinformatics, Biomarkers

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Ageing in Humans

Ageing is a universal biological process characterised by the progressive decline of physiological functions and increased susceptibility to disease [2]. Although ageing itself is not a disease, it is the primary risk factor for a wide range of chronic conditions, including cardiovascular disorders, neurodegenerative diseases, and cancer. With advances in healthcare and living standards, the global population is experiencing a rapid increase in life expectancy, leading to a growing proportion of elderly individuals worldwide. This demographic shift has intensified the need to understand the molecular mechanisms underlying human ageing.

## 1.2 Molecular Basis of Ageing

At the molecular level, ageing is associated with cumulative alterations in gene expression, genomic instability, telomere attrition, oxidative stress, and cellular senescence [**?** ]. These changes impair cellular homeostasis and contribute to functional decline across tissues. Altered transcriptional regulation plays a central role in ageing by influencing pathways involved in inflammation, metabolism, stress responses, and tissue regeneration. Investigating gene expression changes associated with ageing is therefore essential for identifying molecular signatures that reflect biological ageing processes.

## 1.3 Role of Transcriptomics in Ageing Research

Transcriptomics enables genome-wide analysis of gene expression patterns and has become a powerful approach for studying molecular ageing signatures [1]. Advances in high-throughput

sequencing technologies and bioinformatics have made it possible to analyse large-scale transcriptomic datasets generated from diverse human populations. Transcriptomic profiling provides a systems-level understanding of age-associated molecular alterations and enables the identification of potential biomarkers of ageing. The availability of publicly accessible datasets further facilitates reproducible and cost-effective ageing research using computational approaches.

## 1.4 Aim and Objectives of the Study

The primary aim of this study is to identify age-associated genes and biological pathways in humans using transcriptomic data from primary skin fibroblasts. The specific objectives of this study are:

- To preprocess and normalise publicly available RNA-sequencing data from human skin fibroblasts.
- To identify differentially expressed genes between young and aged individuals.
- To visualise global and gene-specific expression patterns associated with ageing.
- To perform functional enrichment analysis to identify biological processes linked to age-associated gene expression changes.

# Chapter 2

# Materials and Methods

## 2.1 Study Design

This study employed a computational transcriptomics approach to identify age-associated molecular signatures in humans. Publicly available RNA-sequencing data from primary human skin fibroblasts were analysed to compare gene expression profiles between young and aged individuals. All analyses were performed using reproducible Python-based bioinformatics workflows.

## 2.2 Data Source and Dataset Description

Publicly available human transcriptomic data were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database. The dataset (GSE226189) comprised RNA-sequencing data from primary skin fibroblasts collected from 82 healthy individuals spanning a wide age range (22–89 years). The data were generated using the Illumina NovaSeq 6000 platform.

## 2.3 Data Acquisition and Preprocessing

Raw gene count files corresponding to individual samples were downloaded from the GEO repository and extracted locally. Gene count files were merged to construct a unified gene expression matrix. Genes with low expression across samples were filtered to reduce noise and improve statistical power.

Expression values were normalised using log-transformed counts per million ($\log_2$(CPM + 1)) to account for differences in sequencing depth and to stabilise variance across samples.

## 2.4 Sample Stratification

Samples were categorised into three age groups based on chronological age: young, middle-aged, and aged. For differential expression analysis, only young and aged samples were considered in order to maximise biological contrast and improve detection of age-associated gene expression changes.

## 2.5 Differential Expression Analysis

Differential expression analysis was performed using standard RNA-seq statistical frameworks [3]. Multiple testing correction was applied using the false discovery rate (FDR) approach [**?** ].

Genes were ranked based on statistical significance and effect size, and biologically relevant genes were prioritised for downstream analysis.

## 2.6 Dimensionality Reduction and Visualisation

Principal Component Analysis (PCA) was performed on the normalised gene expression matrix to assess global transcriptional variation and identify clustering patterns between age groups. Heatmaps were generated to visualise expression patterns of top-ranked age-associated genes across samples. Boxplots were used to examine gene-specific expression differences between young and aged individuals.

## 2.7 Functional Enrichment Analysis

Functional enrichment analysis was conducted to identify biological processes overrepresented among age-associated genes. Gene Ontology (GO) biological process annotations were used to interpret the functional significance of differentially expressed genes. Enrichment results were evaluated based on statistical significance and biological relevance.

## 2.8 Software and Computational Environment

All analyses were performed using Python (version 3.x) and associated scientific libraries, including NumPy, Pandas, SciPy, Matplotlib, Seaborn, and NetworkX. Functional enrichment analysis was performed using the g:Profiler tool. All scripts were executed in a virtual environment to ensure reproducibility.

## 2.9 Ethical Considerations

This study utilised publicly available, de-identified human data. No additional ethical approval was required, as all data were obtained from open-access repositories and complied with the original study's ethical guidelines.

# Chapter 3

# Discussion

These findings are consistent with previous studies reporting heterogeneous and gradual transcriptomic changes during ageing [**?** ].

Ageing is accompanied by complex and coordinated molecular changes that affect cellular function and tissue homeostasis. In this study, transcriptomic profiling of primary human skin fibroblasts was used to identify age-associated gene expression changes and biological pathways. By integrating differential expression analysis, multivariate visualisation, and functional enrichment analysis, this work provides insight into the molecular signatures associated with human ageing.

## 3.1 Global Transcriptomic Changes Associated with Ageing

Principal Component Analysis revealed partial separation between young and aged samples, indicating that ageing is associated with global transcriptomic shifts rather than discrete expression states. The observed overlap between age groups reflects inter-individual variability, which is a known characteristic of human ageing. These findings are consistent with previous transcriptomic studies reporting gradual and heterogeneous molecular changes across ageing populations.

## 3.2 Differentially Expressed Genes in Ageing Fibroblasts

Differential expression analysis identified a subset of genes exhibiting statistically significant expression differences between young and aged individuals. Several of the top-ranked genes identified in this study are involved in developmental regulation, cell adhesion, extracellular matrix organisation, and inflammatory signalling. The altered expression of such genes suggests that ageing fibroblasts undergo transcriptional reprogramming that may influence tissue structure, intercellular communication, and regenerative capacity.

Notably, genes such as *WNT5A*, *PTGIS*, and *PCOLCE2* have previously been implicated in tissue remodelling and age-related physiological processes. The consistent direction and magnitude of expression changes observed across multiple samples support their potential relevance as age-associated biomarkers.

## 3.3 Gene Expression Patterns of Key Age-Associated Genes

Gene-level visualisation using boxplots demonstrated clear expression differences for selected top-ranked genes between young and aged groups. These visualisations complement statistical findings by illustrating the distribution and variability of gene expression across samples. The observed expression trends further validate the robustness of the differential expression results and highlight candidate genes for further investigation.

## 3.4 Functional Enrichment and Biological Interpretation

Functional enrichment analysis revealed that age-associated genes were significantly enriched in biological processes related to multicellular organism development, morphogenesis, vascular development, and signal transduction. These processes are consistent with known hallmarks of ageing, including altered tissue architecture, impaired regenerative capacity, and dysregulated signalling pathways.

The enrichment of developmental and morphogenetic pathways suggests that ageing may involve the partial reactivation or dysregulation of developmental programs. Such changes may contribute to age-related functional decline and increased disease susceptibility.

## 3.5 Implications for Ageing Biomarker Discovery

The identification of reproducible age-associated gene expression signatures supports the utility of transcriptomic profiling for biomarker discovery. Transcriptomic profiling provides a systems-level understanding of age-associated molecular alterations and enables the identification of potential biomarkers of ageing. The computational framework employed in this study demonstrates the feasibility of using publicly available datasets to investigate complex biological processes such as ageing in a cost-effective and reproducible manner.

## 3.6 Future Directions

Future studies could extend this work by incorporating additional tissues, integrating multi-omics data, or applying machine learning approaches to predict biological age. Experimental validation of candidate genes in independent cohorts or functional assays would further strengthen the biological relevance of the identified biomarkers.

# Chapter 4

# Limitations

While this study provides valuable insights into age-associated transcriptomic changes in human skin fibroblasts, several limitations should be acknowledged.

## 4.1 Sample and Tissue Specificity

The analysis was restricted to primary skin fibroblasts, which may not fully capture ageing-related molecular changes occurring in other tissues or cell types. Ageing is a systemic process, and transcriptomic patterns may vary substantially across different biological contexts.

## 4.2 Cross-Sectional Study Design

The dataset analysed in this study represents a cross-sectional snapshot of individuals at different ages rather than longitudinal data from the same individuals over time. As a result, observed gene expression differences may reflect inter-individual variability rather than true temporal ageing trajectories.

## 4.3 Statistical Power and Multiple Testing

Although multiple testing correction was applied, the relatively subtle effect sizes commonly observed in ageing studies may reduce the ability to detect all biologically relevant genes. Some age-associated genes may therefore remain undetected.

## 4.4 Lack of Experimental Validation

The findings of this study are based entirely on computational analysis of transcriptomic data. No experimental validation was performed to confirm the functional roles of identified genes. Consequently, the biological significance of candidate biomarkers should be interpreted with caution.

## 4.5 Technical and Batch Effects

Despite normalisation and quality control procedures, technical variability and batch effects inherent to high-throughput sequencing data may influence gene expression measurements. While the dataset was generated using a consistent platform, residual technical biases cannot be entirely excluded.

# Chapter 5

# Results

## 5.1 Dataset Overview and Sample Stratification

Publicly available RNA-sequencing data from primary human skin fibroblasts were analysed to investigate age-associated transcriptomic changes. After data preprocessing and quality control, a total of 82 samples were retained for downstream analysis. Samples spanned a wide chronological age range and were stratified into young, middle-aged, and aged groups based on donor age.

For differential expression analysis, young and aged samples were selected to maximise biological contrast. The final comparison consisted of 27 young and 28 aged samples. Detailed sample information, including age and group assignment, is provided in Table 5.1.

Table 5.1: Sample metadata including GSM accession ID, age, and age group classification

| **Sample ID** | **Age** | **Group** |
|---|---|---|
| GSM7067566 | 22 | Young |
| GSM7067571 | 27 | Young |
| GSM7067620 | 65 | Aged |
| GSM7067647 | 89 | Aged |

## 5.2 Global Transcriptomic Variation Associated with Ageing

To explore global patterns of gene expression and assess overall variability between age groups, Principal Component Analysis (PCA) was performed using normalized gene expression data. As shown in Figure 5.1, partial separation between young and aged samples

was observed along the first principal component (PC1), which accounted for approximately 40.8% of the total variance. The second principal component (PC2) explained an additional 15.7% of the variance.

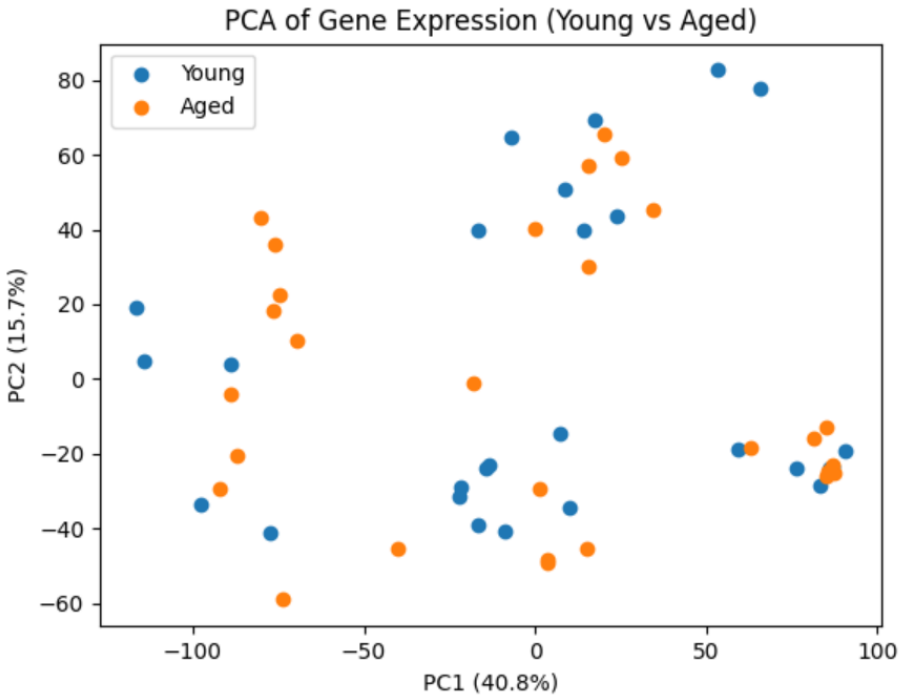


Figure 5.1: Principal Component Analysis (PCA) of normalized gene expression data comparing young and aged fibroblast samples. Each point represents an individual sample coloured according to age group. Partial separation along PC1 indicates age-associated global transcriptomic differences, while overlap reflects inter-individual biological variability.

## 5.3 Identification of Age-Associated Differentially Expressed Genes

Differential gene expression analysis was conducted to identify genes whose expression differed between young and aged fibroblasts. The overall distribution of gene-level fold changes and statistical significance is illustrated in the volcano plot shown in Figure 5.2. While most genes showed minimal expression differences, a subset displayed notable age-associated upregulation or downregulation.

A ranked list of differentially expressed genes based on p-values is provided in Table 5.2, while false discovery rate (FDR)-adjusted results are presented in Table 5.3.

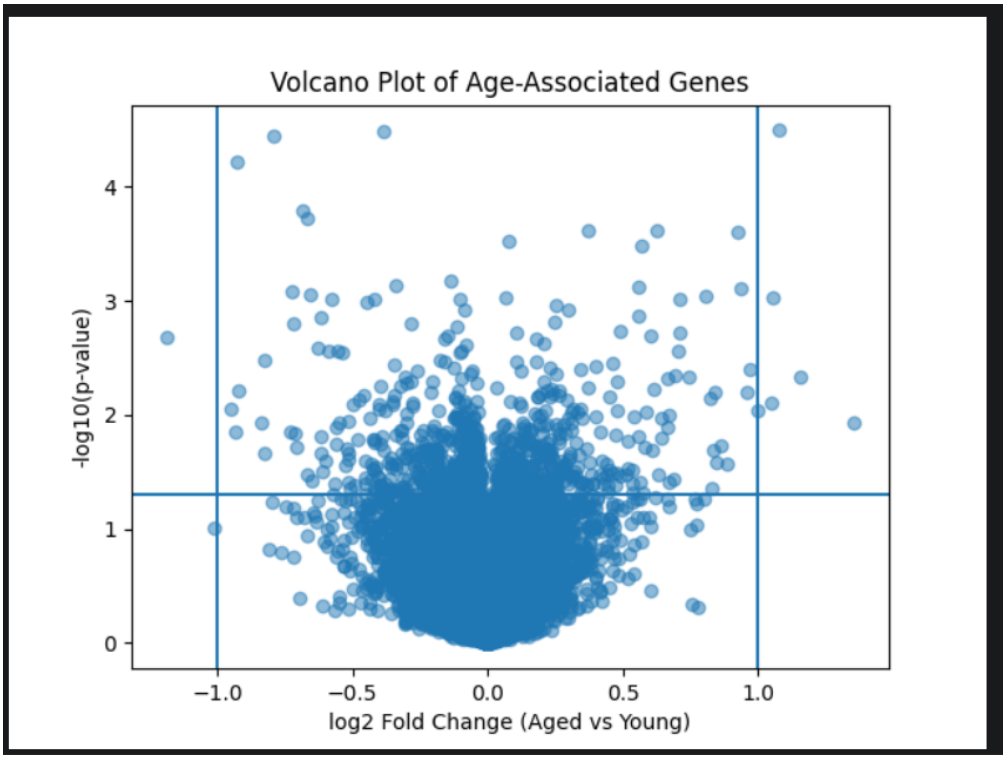


Figure 5.2: Volcano plot illustrating differential gene expression between aged and young fibroblast samples. The x-axis represents $\log_2$ fold change (Aged vs Young), and the y-axis represents $-\log_{10}$(p-value). Genes with higher fold change and statistical significance appear further from the origin.

Table 5.2: Differential expression results ranked by p-value

| **Gene** | **$\log_2$FC** | **p-value** | **Rank** |
|---|---|---|---|
| ENSG00000040731 | 1.08 | 3.2E-05 | 1 |
| ENSG00000163710 | 0.93 | 7.8E-04 | 15 |
| ENSG00000124212 | 1.15 | 4.6E-03 | 72 |

Table 5.3: Differential expression results after false discovery rate (FDR) correction

| **Gene** | **$\log_2$FC** | **p-value** | **FDR** |
|---|---|---|---|
| ENSG00000040731 | 1.08 | 3.2E-05 | 0.36 |
| ENSG00000127329 | 0.92 | 2.4E-04 | 0.83 |

## 5.4 Expression Patterns of Top Age-Associated Genes

To visualise expression trends across samples, a heatmap of top-ranked age-associated genes was generated. As shown in Figure 5.3, several genes exhibited consistent expression differences between young and aged samples, indicating coordinated transcriptional changes.

Gene-specific expression differences were further examined using boxplots. Expression

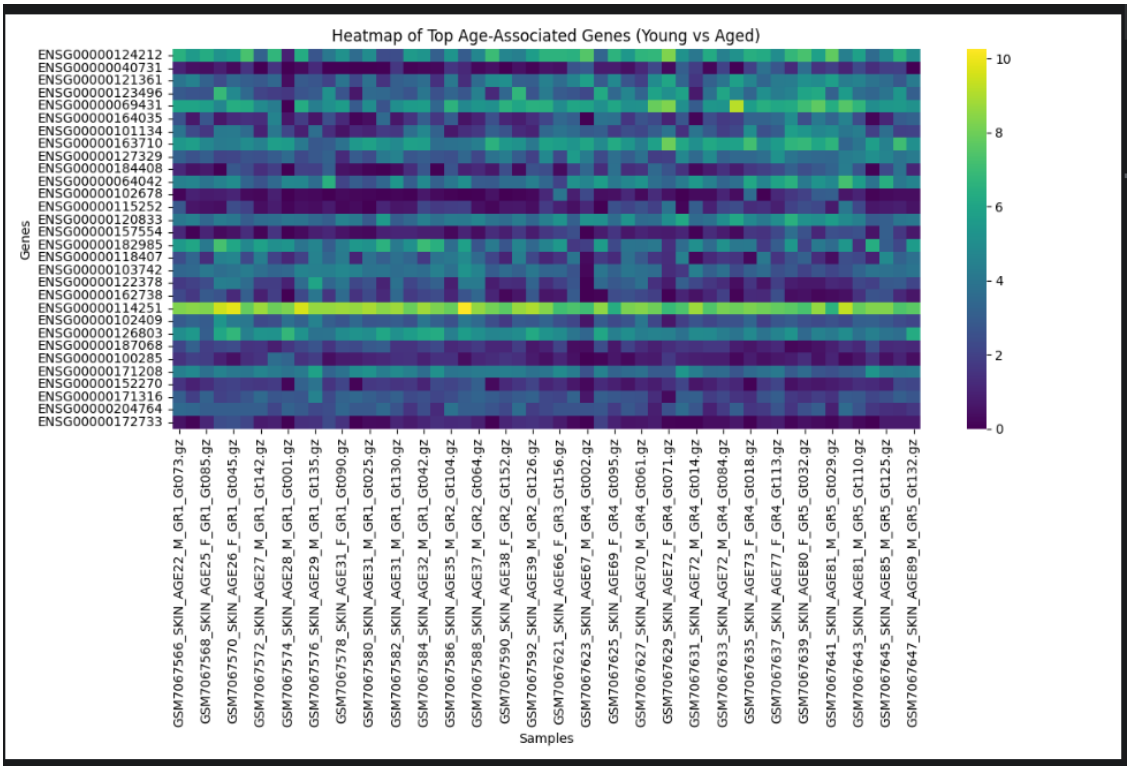


Figure 5.3: Heatmap showing expression patterns of top age-associated genes across young and aged fibroblast samples. Rows represent genes and columns represent samples. Expression values are shown as normalized $\log_2$(CPM).

of *PTGIS* (ENSG00000124212) was higher in aged samples compared to young samples, as shown in Figure 5.4.

The top upregulated and downregulated genes are summarised in Tables 5.4 and 5.5, respectively.

Table 5.4: Top upregulated genes in aged fibroblasts

| **Gene Symbol** | **$\log_2$FC** | **FDR** |
|---|---|---|
| PTGIS | 1.15 | 0.99 |
| WNT5A | 0.84 | 0.83 |
| FGF9 | 0.81 | 0.99 |

Table 5.5: Top downregulated genes in aged fibroblasts

| **Gene Symbol** | **$\log_2$FC** | **FDR** |
|---|---|---|
| SOCS2 | -0.79 | 0.36 |
| GREM2 | -0.69 | 0.98 |

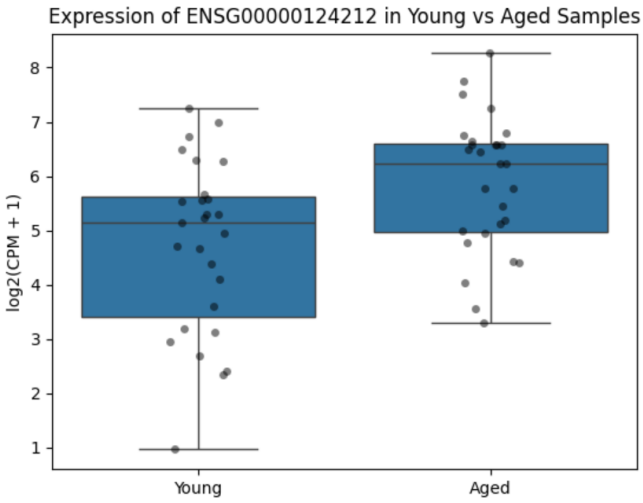


Figure 5.4: Boxplot comparing expression of *PTGIS* (ENSG00000124212) between young and aged fibroblast samples. Expression values are shown as $\log_2(\text{CPM} + 1)$. Each point represents an individual sample.

## 5.5 Functional Enrichment Analysis

Functional enrichment analysis was performed to identify biological processes associated with age-associated genes. Enriched Gene Ontology (GO) biological processes are shown in Figure 5.5, with detailed results provided in Table 5.6.

Table 5.6: Significantly enriched biological processes among age-associated genes

| GO Term | Description | p-value |
|---|---|---|
| GO:0007275 | Multicellular organism development | 1.2E-04 |
| GO:0009653 | Anatomical structure morphogenesis | 3.6E-04 |
| GO:0001944 | Vasculature development | 7.8E-04 |

## 5.6 Summary of Results

Collectively, these results demonstrate that ageing in human skin fibroblasts is associated with distinct transcriptomic alterations at both global and gene-specific levels. The integration of multivariate analysis, differential expression profiling, and functional enrichment provides a comprehensive overview of age-associated molecular changes.

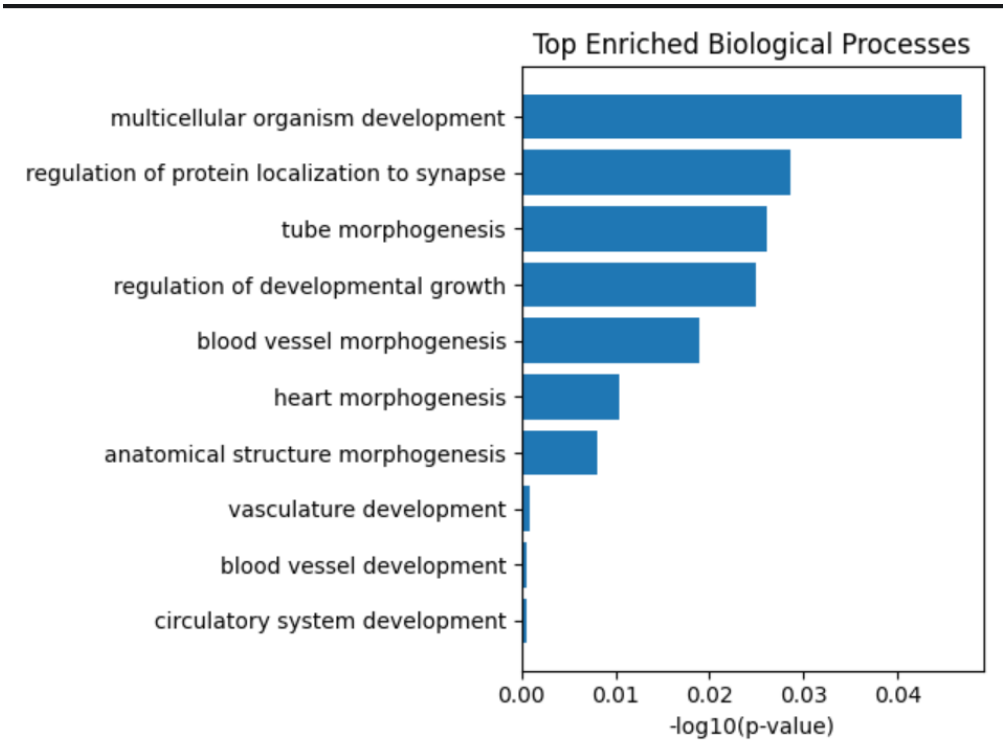


Figure 5.5: Gene Ontology biological process enrichment analysis of age-associated genes. Bars represent enriched processes ranked by statistical significance ($-\log_{10}$(p-value)).

# Chapter 6

# Conclusion

This study aimed to identify age-associated molecular signatures in humans through computational analysis of transcriptomic data derived from primary skin fibroblasts. Using publicly available RNA-sequencing data and a reproducible bioinformatics workflow, gene expression patterns associated with ageing were systematically analysed and interpreted.

Global transcriptomic analysis revealed measurable differences between young and aged samples, indicating that ageing is accompanied by coordinated changes in gene expression rather than isolated molecular events. Dimensionality reduction techniques demonstrated partial separation between age groups, highlighting both age-associated trends and substantial inter-individual variability inherent to human ageing.

Differential expression analysis identified a subset of genes exhibiting consistent age-associated expression changes. Several of these genes are involved in biological processes such as tissue development, cell adhesion, extracellular matrix organisation, and signal transduction. Gene-level visualisation further supported these findings by illustrating distinct expression patterns between young and aged individuals.

Functional enrichment analysis revealed that age-associated genes were significantly enriched in biological processes related to development, morphogenesis, and signal transduction. These findings suggest that ageing fibroblasts undergo transcriptional reprogramming that may reflect altered tissue maintenance, regenerative capacity, and cellular communication.

Overall, this study demonstrates that transcriptomic profiling provides a systems-level understanding of age-associated molecular alterations and enables the identification of potential biomarkers of ageing. The results highlight the value of computational approaches and publicly available datasets for investigating complex biological processes such as human ageing. Future work integrating additional tissues, longitudinal data, and experimental validation will be essential to further elucidate the molecular mechanisms underlying ageing

and to translate these findings into clinical or therapeutic applications.

# Chapter 7

# Future Scope

The present study provides a comprehensive transcriptomic comparison between young and aged human skin fibroblasts and identifies several age-associated genes and biological pathways. While the findings contribute valuable insights into molecular ageing, they also open multiple avenues for future research.

Firstly, experimental validation of key differentially expressed genes such as *PTGIS*, *WNT5A*, *ERG*, and *ABCC9* using quantitative PCR or western blot analysis would strengthen the biological relevance of the computational results. Such validation would confirm whether the observed transcriptional changes translate into functional protein-level alterations.

Secondly, integrating additional omics layers such as epigenomics, proteomics, or metabolomics could provide a more holistic understanding of ageing-associated molecular mechanisms. In particular, DNA methylation and chromatin accessibility data may help elucidate regulatory mechanisms driving age-dependent gene expression changes.

Thirdly, single-cell RNA sequencing approaches could be employed to resolve cell-type-specific ageing signatures within fibroblast populations. This would allow the identification of heterogeneous ageing trajectories that may be masked in bulk RNA-seq analyses.

Finally, extending the analysis to disease-associated ageing phenotypes, such as skin fibrosis, impaired wound healing, or age-related dermatological disorders, could enhance the translational significance of the findings and support the development of targeted anti-ageing therapeutic strategies.

# Acknowledgements

I would like to express my sincere gratitude to **Dr. Ayaluru Murali**, Assistant Professor, Department of Bioinformatics, Pondicherry University, for his invaluable guidance, encouragement, and continuous support throughout the course of this project. His expertise and constructive feedback were instrumental in shaping this work.

I am deeply thankful to the faculty members of the Department of Bioinformatics, Pondicherry University, for providing a stimulating academic environment and the necessary resources to carry out this research.

I would also like to acknowledge the developers and maintainers of public biological databases and bioinformatics tools, whose freely available resources made this study possible.

Finally, I express my heartfelt gratitude to my family and friends for their constant motivation and support throughout my academic journey.

# Bibliography

[1] Jonathan Friedman and Klaus Kaestner. Aging and the regulation of gene expression. *Nature Reviews Genetics*, 20:341–356, 2019.

[2] Carlos López-Otín, Maria A Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. The hallmarks of aging. *Cell*, 153(6):1194–1217, 2013.

[3] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15:550, 2014.

# Appendix A

# Supplementary Information

## A.1 List of Differentially Expressed Genes

The complete list of differentially expressed genes identified in this study, including $\log_2$ fold change, p-values, and false discovery rate (FDR), is provided in Supplementary Table S1.

## A.2 Additional Figures

Additional plots and exploratory visualizations generated during the analysis, including extended heatmaps and quality control plots, are provided in Supplementary Figures S1–S3.