# Unraveling DNA Damage-Induced Ageing Through Transcriptomic Analysis Across DNA Repair-Deficient Models

AI and Biotechnology/Bioinformatics Internship

Authors:
Hasmik Chilingaryan, Swati, James Joy, Temitope Ayano, Gloria

## 1. Abstract

Ageing is a complex biological process driven by cumulative molecular damage, among which DNA damage and impaired repair pathways play a central role. Deficiencies in nucleotide excision repair proteins, such as ERCC1-XPF and CSA, accelerate ageing phenotypes in progeroid mouse models. However, the molecular mechanisms linking DNA repair failure to tissue-specific ageing responses remain incompletely understood. By integrating transcriptomic datasets from DNA-repair-deficient (GSE206778, GSE288213) and naturally aged models (GSE209742), this project aims to identify key genes and pathways involved in DNA damage-induced ageing, offering new insights into genomic instability and longevity regulation.

The primary aim is to investigate how DNA repair deficiency contributes to ageing at the transcriptomic level by identifying differentially expressed genes (DEGs) and pathway alterations. Public RNA-seq datasets are retrieved from the NCBI GEO and ArrayExpress databases. Pre-processing, quality control (QC), and normalisation using DESeq2 or limma are conducted to generate a variance-stabilised expression matrix. Subsequent analysis proceeds in two parallel tracks:

A) DEG analysis and functional enrichment using clusterProfiler and ReactomePA to define molecular signatures, and
B) A Machine Learning framework using a Random Forest classifier to classify experimental groups and identify robust, predictive biomarkers based on feature importance.

Findings from both analytical branches are integrated to prioritise genes consistently associated with genomic instability. This multi-omic approach supports the model that genomic instability is a core driver of ageing and highlights potential biomarkers with clinical relevance.

## 2. Introduction

Ageing is a progressive biological process characterised by a decline in tissue function, reduced cellular resilience, and the accumulation of molecular damage. Among the various forms of age-associated damage, DNA damage and impaired DNA repair mechanisms represent major drivers of cellular and organismal ageing.[1] Endogenous sources such as oxidative stress, replication errors, and transcription-blocking lesions continuously challenge genome stability.[2,3] When repair pathways fail to keep up, cells activate stress responses that alter gene expression, limit proliferation, and contribute to functional decline.

## 3. Research Aim & Objectives

The primary aim of this project is to investigate the transcriptomic mechanisms by which DNA repair deficiency, specifically in ERCC1-XPF and CSA models, accelerates the molecular signatures of ageing, and to identify robust, predictive biomarkers for this process.

### Objectives

The following objectives outline the specific steps taken to achieve the research aim and are mapped directly to the methodological approach.

**Table 1: Specific Objectives and Methods Mapping**

| Objective | Key Analysis | Corresponding Report Section |
|---|---|---|
| 1. Data Foundation | Retrieve and curate RNA-seq datasets, ensuring quality control (QC) and harmonised normalisation. | **Data Retrieval, Preprocessing & QC** |
| 2. Differential Expression | Perform rigorous differential expression analysis (DEG) comparing KO with WT and aged with young samples. | **Track A - Differential Expression & Pathway Analysis** |
| 3. Functional Interpretation | Identify enriched biological pathways (GO, KEGG, Reactome) in DEG lists. | **Track A - Differential Expression & Pathway Analysis** |
| 4. Comparative Analysis | Compare transcriptional signatures derived from the DNA repair deficiency models against the natural ageing dataset. | **Integrated Outputs** |
| 5. Biomarker Discovery | Apply a Machine Learning framework (Random Forest) to classify samples and utilise feature importance to identify potential biomarkers. | **Track B - Machine Learning Analysis** |
| 6. Integration | Integrate findings from DEG, pathway enrichment, and ML to develop a systems-level understanding of molecular ageing. | **Integrated Outputs, Discussion** |

## 4. Dataset Overview

This study integrates three publicly available, high-throughput RNA-sequencing datasets from the NCBI Gene Expression Omnibus (GEO) to investigate the transcriptomic signatures of DNA damage-induced ageing.

**Table 2: Comparative Dataset Table**

| Dataset | GEO ID | Reference | Biological Model | Tissues Analysed | DNA Repair Defect | Samples | Age Points | Platform | Primary Purpose |
|---|---|---|---|---|---|---|---|---|---|
| **Dataset 1** | GSE206778 | [4] | ERCC1 -/-, p53 -/-, WT | HSCs, Hepatocytes, PGCs | NER (Global) ERCC1-XPF | 40 | Young | RNA-seq (Illumina) | Tissue-specific Signatures of ERCC1 Deficiency. |
| **Dataset 2** | GSE288213 | [5] | CSA KO vs WT | Brain Tissue (Cortex, Cerebellum) | TC-NER (CSA protein) | 30 | 45 days, 12 months, 24 months | RNA-seq (Illumina) | Define the Accelerated Ageing Progression signature in the central nervous system. |
| **Dataset 3** | GSE209742 | [6] | WT mice | Various (used for Liver/Brain subset) | Natural Ageing | Varies | Young to old | RNA-seq (Illumina) | Baseline Control for comparison of KO signatures against normal chronological ageing. |

## 5. Methods

### 5.1 Analysis Pipeline Overview

Our analysis pipeline integrates RNA-seq datasets from two DNA repair-deficient mouse models (ERCC1-XPF and CSA knockouts) together with a natural ageing dataset to characterise transcriptomic signatures associated with DNA damage–induced ageing. The workflow begins with dataset retrieval from GEO, followed by preprocessing, quality control, and normalisation to ensure consistency across platforms and conditions. Subsequent analyses proceed along two parallel branches: (1) differential expression and functional pathway enrichment to identify biological processes altered by DNA repair deficiency or ageing, and (2) machine-learning-based biomarker identification using a Random Forest classifier. Outputs from both analytical branches are integrated to determine genes and pathways consistently associated with genomic instability and ageing.

### 5.2 Detailed Methodology

#### 5.2.1 Data Retrieval

RNA-seq datasets were retrieved from the Gene Expression Omnibus (GEO) using the GEOquery package in R. For each dataset, raw or processed count matrices were downloaded depending on availability and suitability for downstream methods. Metadata files were inspected and standardised to ensure consistent annotation of sample characteristics, including genotype, tissue type, and age. All datasets were restructured into harmonised formats to facilitate reproducible analysis across experimental conditions.

#### 5.2.2 Preprocessing & Quality Control

Initial preprocessing involved removing genes with extremely low expression across most samples to reduce statistical noise and improve power. Quality control assessment was performed through principal component analysis (PCA) and sample-to-sample distance heatmaps to detect outliers, batch effects, or inconsistencies in clustering patterns. Where necessary, samples exhibiting abnormal behaviour or conflicting metadata were flagged for review. Group labels (wild-type, knockout, aged, tissue type) were cross-validated with metadata to ensure accurate biological comparisons.

#### 5.2.3 Normalisation

To correct for library size differences and expression variance, normalisation was performed using the DESeq2 framework. Variance stabilising transformation (VST) was applied to produce homoscedastic, log-like expression values suitable for visualisation, clustering, and machine learning. The resulting normalised matrices were used consistently across all downstream computational steps.

Track A - Differential Expression & Pathway Analysis

Differential expression analysis was conducted using DESeq2, with contrasts defined to capture biologically meaningful transcriptional changes, including knockout versus wild-type comparisons and aged versus young comparisons. DESeq2's negative binomial model was used to estimate fold changes, and p-values were adjusted using the Benjamini–Hochberg false discovery rate (FDR) method. Genes meeting significance thresholds were considered differentially expressed. Functional enrichment analysis was performed on DEG lists using clusterProfiler to assess Gene Ontology (GO) Biological Processes, KEGG pathways, and

Reactome signalling networks. These analyses enabled the identification of pathways related to DNA repair, apoptosis, inflammation, metabolic decline, neurodegeneration, and broader ageing processes.

Track B - Machine Learning Analysis
A machine learning pipeline was implemented using the variance-stabilised expression matrix as input for a Random Forest classifier. Data were partitioned into training (70%) and test (30%) sets to evaluate predictive performance. Separate models were trained to distinguish between knockout and wild-type samples and young from aged samples. Model performance was assessed using confusion matrices and classification accuracy metrics. Feature importance scores were extracted to generate ranked lists of genes most influential in distinguishing groups. These machine-learning-identified genes were later compared with DEGs and enriched pathways to prioritise robust biomarker candidates.

### 5.2.4 Integrated Outputs
Findings from both analytical branches were integrated to identify consistent molecular signatures linking DNA repair deficiency to accelerated ageing. Overlapping DEGs, enriched pathways, and machine-learning-ranked genes were examined to determine core biological themes shared between DNA repair-deficient and naturally aged tissues. Tissue-specific patterns were also assessed to evaluate how genomic instability differentially affects stem cells, liver, brain, and germ cell populations. The combined analysis provides a comprehensive multi-omic framework for understanding how impaired DNA repair drives ageing-related molecular decline.

## 6. Results
### 6.1 Differential Expression and Transcriptomic Landscape
Differential expression analysis (DEG) was performed across all three datasets using the DESeq2 framework, revealing a highly context-dependent transcriptomic landscape.

### 6.1.1 ERCC1-XPF Deficiency Analysis
The analysis of the ERCC1-XPF-deficient model (GSE206778 contrast) showed minimal global differential gene expression (Figure 1). Almost all data points clustered tightly along the $\log_2$ Fold Change = 0 vertical line, indicating that few or no genes met the significance threshold for the specific contrast analysed.

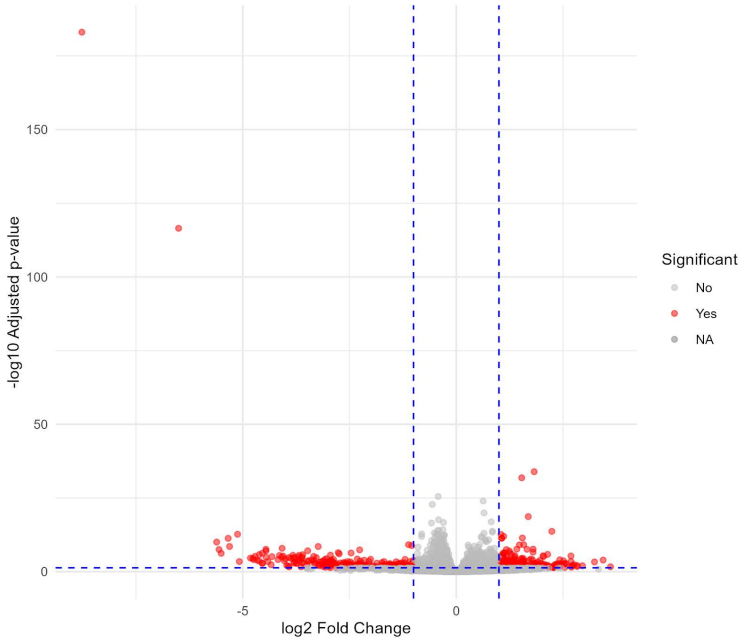


**Figure 1: Volcano Plot of ERCC1-XPF Deficiency (GSE206778 Contrast).** Plot visualising $\log_2$ Fold Change against $-\log_{10}$ (p-value) for the GSE206778 comparison. The extreme vertical clustering indicates a lack of significant differential expression.

### 6.1.2 CSA KO Robust Signature (GSE288213)

Conversely, the CSA KO model (GSE288213) showed a robust transcriptomic change. A clustered heatmap of the top 50 DEGs (Figure 2) demonstrates that genotype (Csa vs WT) is the primary variance driver. A large gene cluster is consistently upregulated (Red) in Csa KO samples across all three age points, demonstrating a robust, sustained molecular signature of TC-NER deficiency.

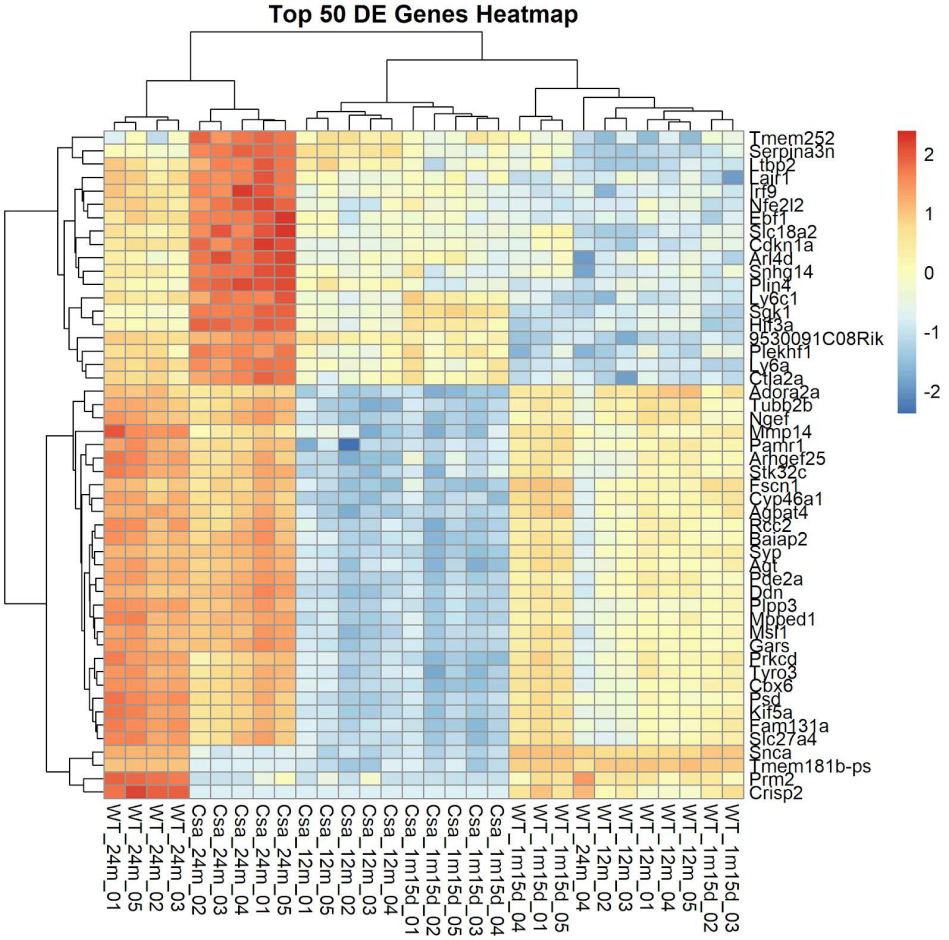


**Figure 2: Heatmap of Top 50 DEGs in CSA KO Brain Tissue (GSE288213).** Clustered heatmap showing normalised expression (Red: High, Blue: Low) of the top 50 DEGs across Csa KO and WT samples at D45, 12M, and 24M. The clustering confirms genotype as the primary expression differentiator.

### 6.1.3 Combined Transcriptomic Overview

The combined view of all datasets confirms a large number of highly significant DEGs in the Natural Ageing and Other Ageing (CSA KO) models (Figure 3). Many points achieved high $-\log_{10}$ (p-value) scores (above 100), providing context for the substantial differences seen between the datasets.

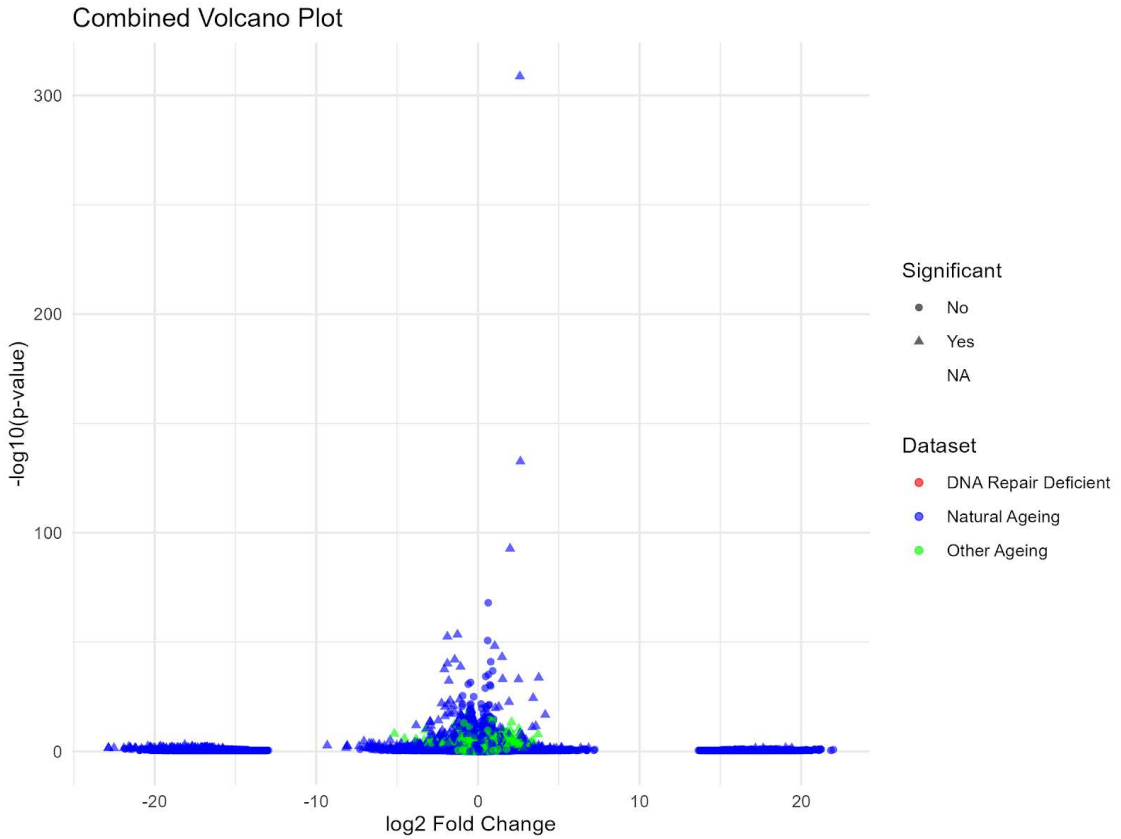


**Figure 3: Combined Volcano Plot of All Datasets.** Plot overlaying DEG results from all three datasets, highlighting the combined landscape of $\log_2$ Fold Change versus significance across the DNA repair deficient, natural ageing, and Other ageing models.

### 6.2 Cross-Dataset Comparative Analysis

To test the core hypothesis of overlapping signatures, a three-way Venn diagram was constructed comparing the DEG lists (Figure 4).

The comparative analysis yielded a critical finding: zero (0) DEGs were shared in the central intersection among the DNA Repair Deficient (GSE206778), Natural Ageing (GSE209742), and Other Ageing (GSE288213) models. This outcome suggests that severe genetic DNA repair defects drive divergent, distinct pathological signatures rather than simply accelerating a common natural ageing pathway.

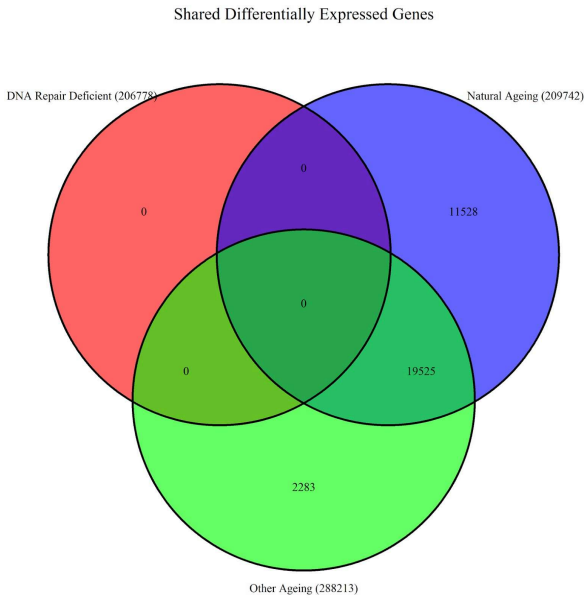


**Figure 4: Three-Way Venn Diagram of Shared DEGs.** The diagram compares DEGs across the DNA Repair Deficient (GSE206778), Natural Ageing (GSE209742), and Other Ageing (GSE288213) datasets. The absence of genes in the central intersection indicates divergence in global transcriptomic signatures.

However, analysis of the two-way intersections showed significant overlap in specific models: a large set of 19,525 DEGs were shared only between the Natural Ageing (GSE209742) and Other Ageing (CSA KO, GSE288213) models. This indicates substantial overlap in some age-related pathways within the CSA model, demonstrating a partial, model-specific resemblance to the natural ageing process.

### 6.3 Functional Enrichment and Pathway Analysis

Functional enrichment analysis was conducted to identify the biological context of the differentially expressed genes (DEGs).

The Combined Pathway Enrichment Dot Plot (Figure 5A) confirmed the core hypothesis that pathways central to the DNA damage response, including DNA Repair, Cell Cycle, and p53 Signalling, are the most significantly perturbed pathways across the DNA damage models.

The specific GO enrichment for the CSA KO model (GSE288213) further highlighted terms related to neuronal function and synaptic transport (e.g., regulation of monoatomic ion transport, exocytosis) (Figures 5B and 5C), which is consistent with the CSA model's known neurological pathology.

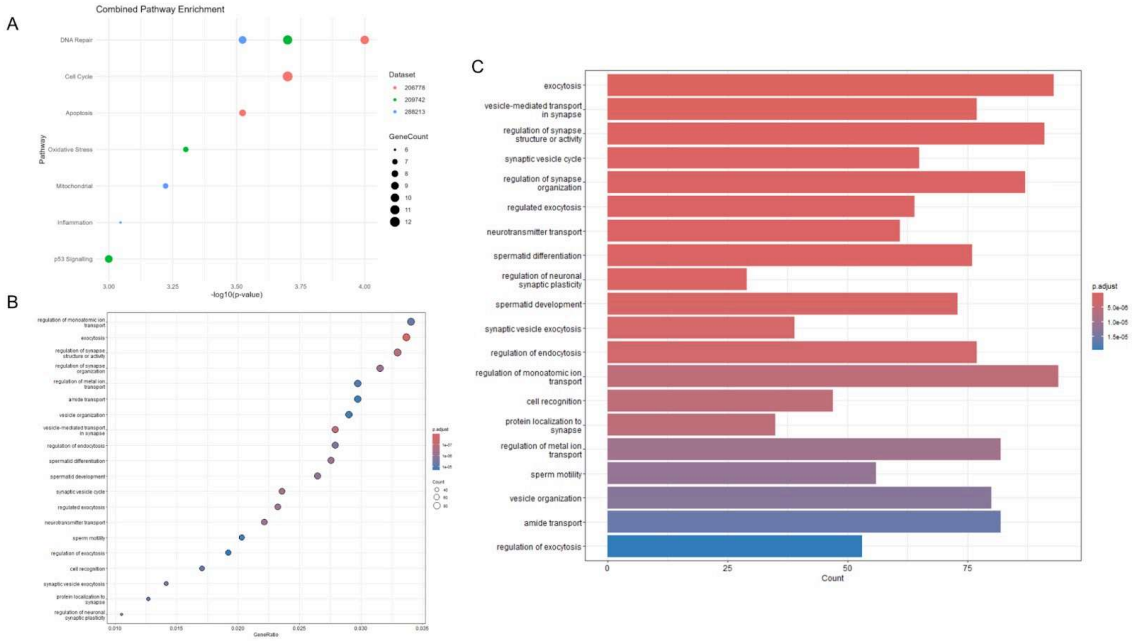


**Figure 5: Functional and Pathway Enrichment Analysis.** This figure presents the functional interpretation of the DEGs across the datasets. **(A)** Combined Pathway Enrichment Dot Plot visualises the significance ($-\log_{10}$(p-value)) and gene count of key a priori ageing and DNA damage pathways across all three datasets. It confirms that DNA Repair, Cell Cycle, and p53 Signalling are the most significantly perturbed pathways in the DNA damage models. **(B)** GO Biological Process Enrichment Dot Plot (GSE288213) and **(C)** GO Biological Process Enrichment Bar Plot (GSE288213) show GO enrichment results for CSA KO DEGs, highlighting terms highly specific to neuronal function and synaptic transport, consistent with the CSA model's neurological pathology.

### 6.4 Machine Learning and Biomarker Discovery

The Random Forest ML model was successfully implemented to classify CSA KO versus WT samples using the GSE288213 transcriptomic data.

The model's feature importance was quantified using the Mean Decrease Gini metric. The top 20 genes ranked by this score are visualised in a bar plot (Figure 6.4.1). Genes such as Zfp365 and Calm1 demonstrated the highest predictive value, positioning them as robust candidate biomarkers for TC-NER deficiency, pending subsequent cross-validation with DEG lists.

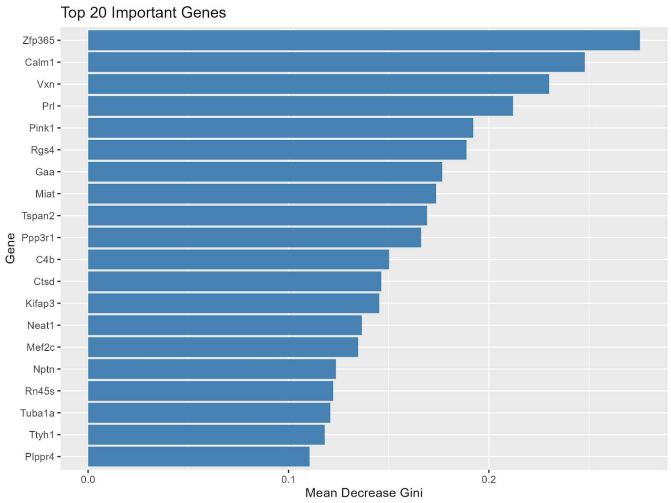


**Figure 6: Random Forest Feature Importance Bar Plot.** Bar plot ranking the top 20 genes by Mean Decrease Gini score, which quantifies the predictive power of each gene in distinguishing CSA KO from WT samples (GSE288213). The top-ranked genes represent the most robust candidate biomarkers.

## 7. Discussion

This study utilised an integrative transcriptomic and machine learning (ML) approach to assess the molecular signatures of DNA repair deficiency in ERCC1-XPF and CSA mouse models and compare them against natural ageing. Our central finding is the divergence of global transcriptomic signatures between the DNA repair-deficient models and the natural ageing dataset. The Venn diagram demonstrated zero shared differentially expressed genes (DEGs) in the central intersection (Figure 4), directly challenging the initial hypothesis that DNA repair defects simply accelerate a common ageing pathway. Instead, severe genetic instability appears to drive distinct, highly specific pathological signatures. The observed transcriptomic changes were model-specific: the CSA KO model (GSE288213) showed a remarkably robust and sustained DEG signature (Figure 2), strongly suggesting the TC-NER defect induces a significant and stable molecular phenotype. This signature clustered with the natural ageing profile in some two-way overlaps, though the global DEG set remained unique. Conversely, the ERCC1-XPF analysis (GSE206778) yielded minimal global DEG (Figure 1), possibly due to analysing only specific stem/progenitor cell populations or rapid compensation in young mice.

Despite the global DEG divergence, pathway analysis confirmed that DNA damage mechanisms remain the core functional driver. The significant enrichment of DNA Repair, Cell Cycle, and p53 Signalling (Figure 5A) across the DNA damage models supports the conceptual model that unresolved DNA damage triggers p53-mediated cellular stress and senescence. The specific enrichment of terms related to neuronal function and synaptic transport in the CSA KO model (Figures 5B, 5C) provides molecular validation for the known neurological pathology of Cockayne Syndrome. Furthermore, the ML framework provided valuable orthogonal validation. The identification of Zfp365 and Calm1 (Figure 6) as top predictive features supports their role as robust molecular markers capable of classifying DNA repair-deficient tissue.

A key limitation is the inherent heterogeneity of the datasets (different tissues, different ages, different genetic defects). Future studies should focus on single-cell RNA-sequencing within a shared tissue across DNA repair models and natural ageing to resolve cell-type specific

DEGs, which may reveal a hidden common ageing signature masked by bulk sequencing. Furthermore, the identified biomarkers require rigorous experimental validation (e.g., using qPCR or Western blot) to confirm their diagnostic utility.

## 7. Conclusion

This integrative transcriptomic study defined the molecular consequences of DNA repair deficiency and tested the hypothesis that genomic instability accelerates natural ageing. We conclude that while DNA damage is a core driver of pathology, it induces divergent, defect-specific molecular signatures rather than simply accelerating a common pathway. Specifically, the CSA KO model (GSE288213) showed a robust, stable transcriptomic signature (Heatmap), consistent with severe TC-NER deficiency, whereas the ERCC1-XPF model (GSE206778) revealed minimal global change (Volcano Plot), suggesting a highly localised response. Crucially, the Venn diagram showed zero (0) DEGs shared in the central intersection, indicating that genetic repair defects drive distinct pathological processes. Nonetheless, pathway analysis confirmed that the underlying mechanism involves central DNA damage response pathways, DNA Repair, Cell Cycle, and p53 Signalling, supporting the model that unresolved damage triggers cellular stress and functional decline. Finally, the Random Forest ML model successfully identified highly predictive candidate biomarkers, such as Zfp365 and Calm1, which serve as robust molecular markers to distinguish DNA repair-deficient tissue and warrant further experimental validation.

## References


1. Dataset 1 Paper Link: https://pubmed.ncbi.nlm.nih.gov/38514641/
2. Dataset 2 Paper Link: https://pubmed.ncbi.nlm.nih.gov/40570619/
3. Dataset 3 Paper Link: https://pubmed.ncbi.nlm.nih.gov/37348497/
4. Background Figure Article: https://www.sciencedirect.com/science/article/pii/S0092867405001029#fig6